DUDLEY KNOX LIBRARY
NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA 93943

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

A USER-ORIENTED
PERFORMANCE INDEX FOR
PACKET SWITCHED NETWORKS

by

Mark E. Speck

March 1986

Thesis Advisor: J. W. LaPatra

Approved for public release; distribution is unlimited.

T227285

RITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

REPORT SECURITY CLASSIFICATION NCLASSIFIED	1b. RESTRICTIVE MARKINGS
SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
DECLASSIFICATION / DOWNGRADING SCHEDULE	
ERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S)
NAME OF PERFORMING ORGANIZATION aval Postgraduate School	6b. OFFICE SYMBOL (If applicable) 32
7a. NAME OF MONITORING ORGANIZATION Naval Postgraduate School	
ADDRESS (City, State, and ZIP Code) onterey, California 93943-5100	7b. ADDRESS (City, State, and ZIP Code) Monterey, California 93943-5100
NAME OF FUNDING / SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)
9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER	
ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.

TITLE (Include Security Classification)
USER-ORIENTED PERFORMANCE INDEX FOR PACKET SWITCHED NETWORKS

PERSONAL AUTHOR(S)
peck, Mark E.

TYPE OF REPORT	13b TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
aster's Thesis	FROM _____ TO _____	1986 March	77

SUPPLEMENTARY NOTATION

COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Networks, packet switched networks, performance, measures of performance

ABSTRACT (Continue on reverse if necessary and identify by block number)

As packet switched networks become more widely used, there will be a need to characterize network performance with an index that defines how well the network can meet the user's needs. This thesis explains present and future measures of network performance, some of which are user-oriented, and how they can be useful. Dividing users into groups based on user needs, a model of a composite performance index is developed. The usefulness of this index, if provided on a real-time basis, is then explored and substantiated.

0. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION unclassified
2a. NAME OF RESPONSIBLE INDIVIDUAL J. W. LaPatra	22b. TELEPHONE (Include Area Code) 408-646-2249
	22c. OFFICE SYMBOL 54Lp

D FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Approved for public release; distribution is unlimited.

A User-Oriented Performance Index for Packet Switched Networks

by

Mark E. Speck
Lieutenant, United States Navy
B.S., St. Mary's College, 1977

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN TELECOMMUNICATIONS SYSTEMS MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
March 1986

ABSTRACT

As packet switched networks become more widely used, there will be a need to characterize network performance with an index that defines how well the network can meet the user's needs. This thesis explains present and future measures of network performance, some of which are user-oriented, and how they can be useful. Dividing users into groups based on user needs, a model of a composite performance index is developed. The usefulness of this index, if provided on a real-time basis, is then explored and substantiated.

Thesis
S6672
c.1

TABLE OF CONTENTS

I. INTRODUCTION 9

II. PACKET SWITCHED NETWORKS 12
A. INTRODUCTION 12
1. Circuit Switching 12
2. Message Switching 13
3. Packet Switching 13
B. HISTORICAL DEVELOPMENT OF PACKET SWITCHED NETWORKS 14
C. PACKET SWITCHED NETWORK BASICS 16
1. Background 16
2. Packet Structure 17
3. Routing Techniques 19
4. Virtual Circuits and Datagrams 20
5. Protocols 22
6. Flow and Congestion Control 24
D. NETWORK MANAGEMENT 27
1. Control Functions 29
2. The Network Monitoring Center 29
3. Summary 30

III. MEASURING NETWORK PERFORMANCE 32
A. INTRODUCTION 32
B. HOW SHOULD WE MEASURE ? 32
C. WHAT DO WE MEASURE 34
1. Throughput 35
2. Delay 37
3. Error 38
4. Other Parameters 40
D. STANDARD NETWORK PERFORMANCE PARAMETERS . . . 42

IV. A USER-ORIENTED PERFORMANCE INDEX 48
A. BACKGROUND 48
B. PERFORMANCE REQUIREMENTS FOR VARIOUS USER TYPES 50
C. CHARACTERISTICS OF A USER ORIENTED, REAL-TIME PERFORMANCE INDEX 51
D. THE MODEL FOR ESTIMATING A NETWORK PERFORMANCE INDEX 53
1. Throughput (Rate) 54
2. Delay 56
3. Quality 59
4. A Composite Performance Index 60
5. Stability 62

V. THE FUTURE 64
A. CONCLUSIONS 67
B. AREAS FOR FURTHER RESEARCH 69

LIST OF REFERENCES 73

INITIAL DISTRIBUTION LIST 76

LIST OF TABLES

I	PACKET SWITCHED NETWORK ROUTING REQUIREMENTS . .	20
II	ISO 7-LAYER PROTOCOL	23
III	FLOW AND CONGESTION CONTROL CRITERIA	27
IV	ROLE OF NETWORK CONTROL IN A PACKET SWITCHED NETWORK	29

LIST OF FIGURES

2.1 Basic Packet Switched Network 16
2.2 Structure of a Typical Packet 17
3.1 Summary of ANSI X3.102 Performance Parameters . . . 45
4.1 Utility Versus Throughput 55
4.2 Possible Utility Profile 57
4.3 Utility Versus Delay 58
4.4 Utility Versus Quality 60

ACKNOWLEDGEMENTS

This research was supported, in part, by the National Communications System.

I. INTRODUCTION

Presently our world is evolving in a direction resulting in the generation of tremendous amounts of information. This information is often of little value when it is restricted to individuals, that is to say, there are usually benefits to be derived from sharing information. Exchanging ideas, avoiding duplication of research effort, and the ability to utilize the facilities of others are only a few examples of the benefits that information sharing can provide.

This information exchange is done by using networks: social networks such as conferences and meetings, written media such as books and papers, and now with new technology much of this information exchange is being done using telecommunication networks. A particular type of network receiving increased notoriety due to its speed and efficiency is the packet switched network.

The rapid growth of technology and hardware has led to the development of networks capable of tremendous amounts of information exchange. These networks are becoming more widely used by the public sector. The network capabilities have outpaced our ability to manage and control these networks in that we are constantly finding ways of extracting more capability out of existing networks in terms of what it can provide to the user. With managers becoming overwhelmed at what technology can accomplish, the user is doubly overwhelmed.

In utilizing the service provided by a network, the user developes an impression of how well the network is meeting his needs - he is in essense developing a subjectively determined network performance level at the particular instance in time. The actual performance level is changing constantly, usually at a rate faster than the user is

capable of subjectively determining. An objective measure of network performance based upon actual network status, if available to the user, could be very valuable. It could likely enhance the user's benefit from the network in that he would know from the start of his network session how well the network is performing. This could result in more optimal use of his time, money, and communication resources, not to mention reducing the user's aggravation in using the network. It could also be of benefit to the network itself in that it may more evenly spread the demand on an already overloaded network.

Many things affect how a network performs. The speed and accuracy of the hardware, the efficiency of the software, as well as how these various resources are linked together determine the maximum performance level of the network. The users demands and how the network managers and operators control the network have a significant impact on how the network is meeting the needs of the user at particular time. This thesis will investigate how a user-oriented performance measure could benefit both network users and managers.

A fundamental question is" What network parameters do we measure to determine its performance?" Numerous technical details of the inner workings of the network are measured and are extremely useful to the manager in performing his task of planning, resource allocation, configuration control, and status monitoring all of which hopefully result in optimization of network performance. Significant standardization of what to measure and how to measure it does not exist and that which is measured is of little value to the user in that it is either too technical, too multi-faceted, or not timely.

Standardized performance parameters are being investigated by several organizations such as the American National Standards Institute (ANSI), the National Telecommunications

and Information Administration (NTIA), and the International Telephone and Telegraph Consultive Committee (CCITT). Some strides have been taken towards developing user-oriented performance parameters but only to the extent that they recognize that network performance should be optimized with the user in mind. This applies to network planning, design, and management control. This is a step in the right direction but falls short of developing a measure of network performance that is of value to the user at the time he actually wishes to use the network.

This thesis, written with the network user in mind, will first explain what packet switched networks are, how they work, and why they are becoming more widely used in the next chapter. Chapter three explains how packet switched network performance is presently measured and how they are used. A method of determining a user oriented performance figure-of-merit for a packet switched network is explained in chapter four with three classes of users in mind; large volume data users, interactive processing users, and voice users. Chapter five outlines future developments in network performance monitoring, conclusions, and areas for further research.

II. PACKET SWITCHED NETWORKS

A. INTRODUCTION

Communication networks can be divided into several different types but probably the most common network type classification is [Ref. 1] :

1. Circuit-switched
2. Message-switched
3. Packet-switched

1. Circuit Switching

In circuit switching, a "call" is set up between users and message transmission takes place in a conversational mode. To set up the call, a path of connecting transmission lines must be established between source and destination. When a user wishes to communicate with someone, a signal is transmitted through the network, seizing available channels in an attempt to reach the destination. If no path is available, a signal is sent back to the caller indicating he must wait until the necessary channels are free. When a path has been established, message transmission takes place over all of the channels in the path in both directions, simultaneously. The path remains allocated to the users until the caller releases the path or until the call is preempted by a user of higher priority.

For many years, circuit switching was the most effective means of communicating and is still considered the most effective means of accomodating voice traffic but can also be used for data transmission. Examples of circuit switched networks include the public telephone network and the Department of Defense (DoD) Automatic Voice Network (Autovon).

2. Message Switching

In message switching, transmission of data takes place in what is known as "store and forward" procedure. Unlike circuit switching, only one channel is used at a time for a given message transmission. The message proceeds from the source to another node in the network according to a predetermined route depending on source and final destination. When the message arrives at a node it is stored. It may be forwarded immediately to the next node in its path, or if no channels are available, it is placed in a queue of messages awaiting transmission and forwarded when the channel becomes available. The message proceeds through the network in this manner until it arrives at its destination. This way, the message only needs to occupy one channel at a time rather than the entire path from source to destination. This frees up channels for additional message transmission. The DoD Automatic Digital Network (Autodin) is an example of a message switching network.

3. Packet Switching

Packet switching is very similar to message switching in that it relies on the store and forward technique for message transmission. A major difference is that the messages are broken down at the source into smaller parts known as packets. These individual packets are then transmitted through the network in a manner similar to messages in a message switched network. When all of the packets reach the final destination, they are reassembled back into the form of the original message and it is sent to the individual user. The individual packets of a message may take very different routes in reaching the destination allowing them to avoid long delays at a node. With the possibility that the packets may follow different routes simultaneously, significant savings in message delay can be realized. This also leads to the possibility of out of

sequence arrival at the destination. To remedy this, each packet has some identifier as to what message it belongs to and in what order the packets are to be reassembled at the destination. Specific attributes of packet switched networks will be explained more fully later in the chapter.

B. HISTORICAL DEVELOPMENT OF PACKET SWITCHED NETWORKS

Packet switching technology was not really an invention but rather a reapplication of basic dynamic allocation techniques used for over a century by telegraph and torn paper tape switching systems [Ref. 2: p. 1307]. These techniques had traditionally used manual sorting and routing decisions and were therefore considered to be less efficient than circuit switching techniques. In order for packet switching techniques to show the necessary efficiency, both computer processing power and buffer storage resources are needed at each network node. The resulting economic tradeoff was simple: if lines are cheap, use circuit switching; if computing is cheap, use packet switching. With significant drops in the price of computing power, packet switching has become the economical choice.

The first published description of what is now known as packet switching was an eleven volume analysis written by Paul Baran of the RAND Corporation [Ref. 3] in 1964. This report showed significant advantages in the use of packet switching in terms of speed and cost but sat largely ignored for several years. Independently, but almost concurrently, the Advanced Research Projects Agency (ARPA) was interested in linking time-shared computer facilities together through a widespread communications network. Soon the details of a store-and-forward switched network similar to what was proposed by Baran were conceived [Ref. 2: p. 1308]. It was later that the two developing groups learned of the other's efforts.

Even though significant economic advantages were expected with packet switching, the communications world was hard to convince. In 1967 ARPA published a plan to link computer facilities together. Minicomputers, used as dedicated switches, were to be interfaced with the computing facilities and interconnected by leased lines. In 1968, Bolt, Beranek, and Newman, Inc. (BBN), was awarded a contract to develop the network. In December, 1969, a four node network was installed and operating. The network expanded rapidly and by March of 1977, 111 nodes were operating, linking thousands of facilities and users involved in various research and development programs.

The Department of Defense (DoD) recognized the potential benefits of using this ARPANET technology and in 1982 the Deputy Secretary of Defense ordered the termination of a program to be used for DoD data communications (AUTODIN II) [Ref. 4]. Instead, service heads were to proceed with the implementation of the Defense Data Network (DDN) which would utilize ARPANET technology and would integrate all DoD data communication users into a common network. Presently the DDN consists of several networks devoted to particular groups of users with the capability of interconnecting users of different networks by using what are commonly called 'gateway' nodes. The network is continuing to expand and presently consists of hundreds of nodes and covers all of the United States, Japan, Korea, the Philipines, as well as several NATO countries [Ref. 5: p. 3,4]. The ARPANET still exists as a subnetwork of the DDN, devoted to research and development.

Several countries have developed government and publicly sponsored packet switched networks including France, Canada, Australia, and the United Kingdom [Ref. 2: p. 1309] and their number and geographic coverage continues to expand.

C. PACKET SWITCHED NETWORK BASICS

1. Background

Figure 2.1 illustrates the components of a basic packet switched network. The "user terminal" could have a variety of forms such as a teletype keyboard and printer, an interactive data and graphics CRT terminal, a telephone handset and associated digital interface, or a software process in a computer. It could even represent a whole independent local network joining multiple user facilities in an office or school campus.

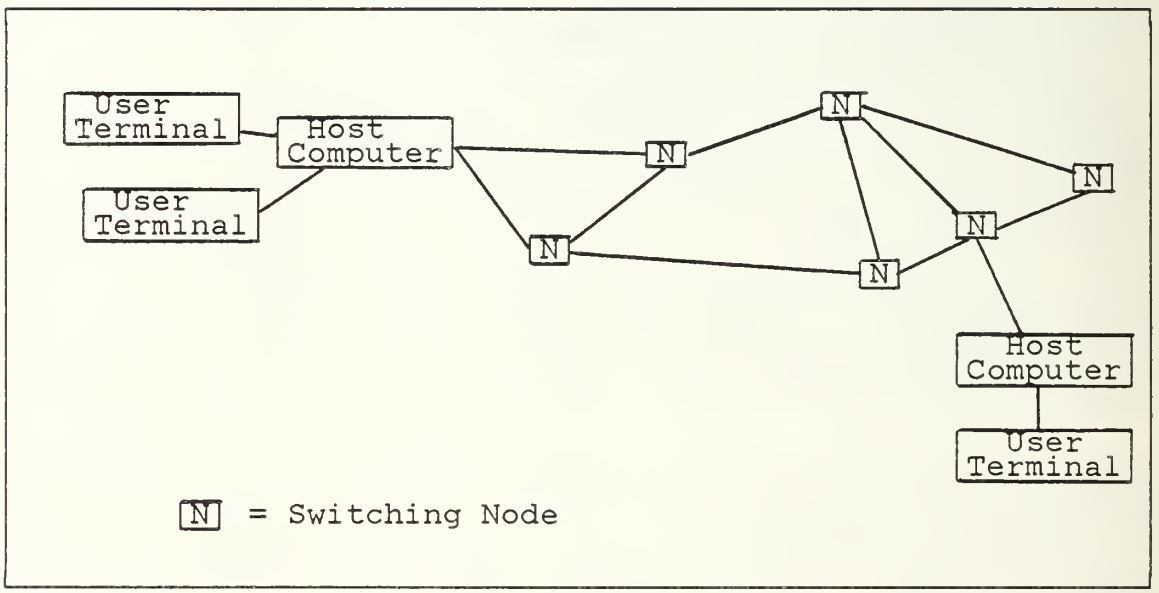

Figure 2.1 Basic Packet Switched Network.

The "host computer" in Figure 2.1 provides the user terminals attached to it, all of the services and translations needed to get their data to and from the network. The host will provide packetizing and depacketizing service for "dumb" terminals and lesser services for "intelligent" terminals or networks.

The host computers are joined to network nodes via some type of transmission media. These nodes have relatively powerful switching and routing control computers which

receive, sort, store, and forward packets to the network using route selection algorithms which accomodate network conditions and performance objectives. These nodes are linked to several other nodes with various types of transmission media. Note that there are several different routes for data to flow between the two users depicted.

The network may utilize one or several different routes during a network session in order to efficiently use the network resources. Here lies one of the most powerful attributes of packet switched networks.

Each packet received at an intermediate node is typically stored in a buffer queue until processing capacity is available to decide what to do with the packet. When the packet is at the head of the queue, the processor examines the packet for its destination address and decides what node to send it to next and places the packet in a queue awaiting transmission via a particular link. Each of these operations takes time, thus contributing to the end-to-end delay for the packet or packets to traverse the network.

2. Packet Structure

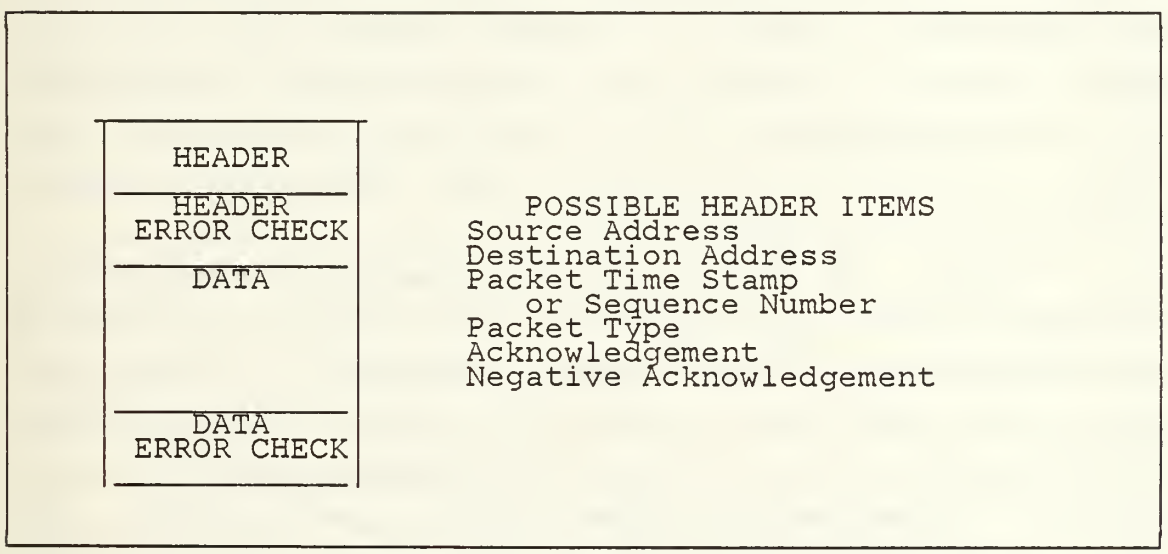

Figure 2.2 Structure of a Typical Packet.

Figure 2.2 illustrates a typical packet format as constructed by a host computer. After the digital information source associated with the user terminal has produced enough bits to fill the "data" bits field, the host attaches a carefully formatted block of header items. The error check field typically contains a "checksum" generated by Cyclic Redundancy Check (CRC) hardware/software at the computer connected to the originating node. This CRC characterizes all the information according to a specified pattern. Intermediate or destination switching computers may calculate the same checksum on the received bits and if any disagreement exists between calculated and received error check field then it can be assumed that bit errors occurred in transmission. The action that the node takes will depend on the nature of the data. Some data, such as a computer file or a banking transaction, may require to be error free. In this case the node will request the sending node to retransmit the packet. This is normally done by sending a negative acknowledgement (NACK) or if no acknowledgement is received by the sending node after a certain time period, the packet will be retransmitted. Packet switched speech, on the other hand, may not be significantly affected by a modest number of errors so any errors detected in the data can usually be ignored. Errors occurring in the header must not be ignored as it may create problems in packet recontruction.

All of the necessary packet switching transactions within and between terminal hosts and networks are governed by defined sets of rules called protocols. In essence, protocols provide the participants with a standardized vocabulary of message exchanges and responses needed to: 1) initiate a communication session with a destination; 2) request and receive transmission capacity; 3) format, transmit, and acknowledge packets; and 4) terminate the

session. A well designed protocol enables packet switched communications between completely different host computers in much the same way that a high level computer programming language such as Fortran or Pascal enables software systems to run on dissimilar machines. More on protocols later.

3. Routing Techniques

In a large, highly connected packet network there can be many possible paths between a source and destination. Tremendous amounts of research has been devoted to finding the "best" path between two points in a network, where "best" can be defined as minimum length, minimum delay, least cost, least number of hops, maximum flow rate or some combination of these. Each node has a degree of independence in choosing the next node in order to achieve the best path in the end-to-end sense. Various adaptive techniques have been used to permit nodal processors to account for large scale network conditions such as regions of temporary congestion or link outages in order to try to improve the end-to-end quality of paths.

Often a fixed or static routing table is used where each source-destination pair is evaluated and a best path is developed and placed in a table. In this way, when a destination node notes a packet's destination, it merely checks the table and sends the packet to the designated node. This is simple but due to the growth of large networks, changes to links and nodes occurring more frequently, and instances of temporary congestion, the result is frequent poor performance in packet delivery [Ref. 6: p. 115]. Because of this, almost all long-haul packet switched networks use some sort of dynamic routing algorithm which bases the routing decision on measurements or estimates of the current traffic and topology of the network. Table I [Ref. 7: p. 12] lists recognized requirements for Packet Switched Networks.

TABLE I

PACKET SWITCHED NETWORK ROUTING REQUIREMENTS

Message routing should ensure rapid and error free packet delivery

Routing strategy should adapt to changes in network topology resulting from node and communication link failures

Routing techniques should adapt to varying source/destination traffic loads

Packets should be routed around congested or blocked nodes

Packets should be routed to their destination via some least cost method

A technique for detecting looping and 'ping-ponging' should be included

Routing techniques should be as simple as possible to minimize hardware and software requirements

The DDN uses a routing strategy known as adaptive directory source routing [Ref. 6: p. 121]. Each node sends a packet to all other nodes every ten seconds called a Link State Packet [Ref. 8: p. 93]. This packet contains information about the state of a node's links with all its neighbor nodes in terms of delay for a packet to be sent along that link. A simple but elegant broadcasting technique known as intelligent flooding ensures that all nodes receive the Link Status Packet. Upon receipt of each new Link Status Packet, the node develops a new routing table for sending normal packets traversing the network. This table generation requires an efficient but complete database that enables each node to essentially have a complete map of the network. It also must attempt to maintain consistent network mapping from node to node [Ref. 9].

4. Virtual Circuits and Datagrams

There are two basic kinds of service that a packet network can provide in terms of how it delivers packets. In datagram service each packet is treated as an independent

entity known as a "datagram". Each datagram is routed and delivered independently even though a message from a source to a destination may consist of several packets. The use of this scheme requires each packet to contain all the information necessary for its routing including source, destination, message identifier, a sequence number, and any other information required by the routing scheme. Because of this, datagram service has a high probability of delivery but not an absolute probability. There is a chance that one can get lost with the destination having no knowledge that it had ever been sent. Also the message reconstruction requires that all packets be ordered properly before message delivery can occur.

By contrast, the virtual circuit is a sequential service, that is, the packets arrive at the destination in the same order in which they entered the network. This is guaranteed by requiring that all of the packets follow the same route as the first packet. This is accomplished by a circuit set-up. The originating host sets up buffer space along an entire path to the destination. When the destination acknowledges that it is ready to receive the message, the packets are sent in sequential order. In this case, each packet needs only to identify to what virtual circuit it belongs to and the different switching nodes will automatically route it through the links that correspond to its virtual circuit. After the last packet, a circuit termination procedure occurs which frees up the resources for other use. A virtual circuit requires less overhead bits per packet but at the expense of call set-up and termination times. It is absolutely reliable in packet arrival but lacks some flexibility in routing should a failure occur during a virtual circuit session. Another benefit is that it reduces routing burden at each switching node.

A virtual circuit resembles the circuit switching technique in the sense that the circuit is set up and terminated and that all packets of a message traverse the same route but differs from it significantly in that more than one virtual circuit may simultaneously share the same link as part of its message path. Another important difference is that if a link or node does fail, alternate routing can occur resulting in delay but not in lost communications as would occur in the case of circuit switching. To the user, a virtual circuit appears to be a dedicated line.

Two advantages frequently given to to datagram service are improved routing reliability and lower cost. Due to advanced switching and storage capabilities at nodes, link or node failures no longer result in lost packets and the packet retransmissions they cause in the case of the virtual circuit. Drops in memory and switching costs result in the extra overhead associated with datagram service becoming a major cost factor. A datagram requires about 25 bytes of header in addition to the actual data (0 to about 128 bytes) whereas only about 8 bytes are required for similar packets in a virtual circuit. Messages requiring more than one packet cause overhead to contribute 13 to 94% to the total transmission costs in the case of datagrams [Ref. 2: p. 1311].

5. Protocols

The design, development, and standardization of packet system protocols has received a great deal of attention. Protocols exist to ensure that communicating entities can send, receive, and interpret the information that they wish to exchange. Protocols play three fundamental roles: 1) they establish standard data elements; 2) they establsh conventions; and 3) they establish standard communication paths. Data elements include character, messages, files, jobs, and graphic displays. Conventions include code sets,

packet formats, transmission speeds, and control messages. Standard communication paths can include functions such as addressing, priorities, sequencing, error control, flow control, and session initiation and termination.

Two protocols in extensive use today include the International Standards Organization (ISO) Open System Interconnection 7-layer architecture and the CCITT X.25 protocol. In the ISO model, the highest layers govern end-to-end interaction between users or internal processes and the lowest layers govern the operation of the media carrying the actual signals on the individual links. Intermediate protocol layers apply to switching nodes and intermediate processes. The 7-layer architecture is designed for virtual circuit service but work has been done to include datagram service. Table II shows the 7 layers. For a more in-depth explanation of each of the layers, see Stallings [Ref. 18].

TABLE II
ISO 7-LAYER PROTOCOL

7. Application Layer
6. Presentation Layer
5. Session Layer
4. Transport Layer
3. Network Layer
2. Data Link Layer
1. Physical Layer

The CCITT X.25 protocol was first agreed upon internationally in March 1976 and like other standards, continues to evolve [Ref. 2: p.1310]. It is a three layer protocol with the layers closely corresponding to the 3 lowest layers of the ISO protocol (physical, data link, and network). The layers are named physical layer, data link layer, and packet layer. Higher layers are provided by the user's application

software. X.25 is a virtual circuit service. For a more in-depth look at X.25 see Rosner [Ref. 11].

The DDN uses a protocol similar to those above, but because it is designed to work in an internet environment (where source and destination reside on different networks), the provision of error-free, end-to-end transmission becomes more complicated. The DDN uses internet standards called Internet Protocol (IP) and Transmission Control Protocol (TCP) which operate at the user level. Error checking is included such that end user processes can detect errors that were introduced at some (possibly unidentifiable) intermediate node and request retransmission by the source.

Protocols are necessary to allow user-to-user communication and thus are highly desireable. Standardization is also desireable because of widespread differences in users and their needs as well as the existence of many organizations involved in standard development. It is often the case where "universal" standards or families of standards are developed and adhered to by some and ignored by others.

6. Flow and Congestion Control

Certain similarities exist between packet switched networks and traffic on major highways. A large number of users (cars or packets) share the resources of the network. Unless careful control is exercised over user demand, capacity may be exceeded. During rush hour, the number of cars often exceeds highway capacity creating long lines at highway interchanges. Interference between "through" traffic and traffic using on and off ramps reduces the throughput of the highway. If the situation persists, traffic may come to a standstill. This is a good example of the relationship between offered load and throughput. After reaching system capacity, if offered load is increased, throughput drops off. Some means of controlling the offered load to prevent the system from exceeding the overload point is needed.

Just as highways use access ramp control lights to limit flow onto the highway, packet switched networks need some similar controls. These are frequently called flow and congestion control. There are some subtle differences between flow and congestion control. Flow control is principally involved with controlling the rate at which packets flow to prevent their arrival rate from exceeding the rate that a processor can handle them. Flow control can be defined as a system of algorithms used in a network to prevent a user or group of users from monopolizing resources to the detriment of others.

Flow control can take place between neighboring nodes, end-to-end users, or source to destination switches. Three types of flow control are generally utilized in networks [Ref. 12: p.107] :

1. Rate: The source is restricted to sending only up to N packets per second.
2. Window or Credit: The sending node may send up to W packets before receiving an acknowledgement. No more than W unacknowledged packets may exist between source and destination.
3. Stop and Go: A source sends packets at its maximum rate, stops, then repeats. This can be controlled by a timer or more commonly, by a feedback packet from a downstream node or destination.

Flow control methods can be either or dynamic. In static control, some constant level of restraint exists in the network at all times. This is simple, but overly restrictive in terms of network throughput under light loads, and conversely, may be insufficient during peak periods of many users. Dynamic controls apply variable levels of flow control (e.g. changes in window size), depending on network conditions. This variability can either be in response to changes in load or can be anticipatory estimates of short-term future load. The network responds by inducing flow restriction before the increased load degrades the network. The ideal response of a flow control system would be to immediately reduce traffic load

to the maximum manageable level upon sensing an overload limit being reached. This ideal is not achievable due to the fact that a finite time is needed for flow control actions to take effect. If flow control actions are too strong, large load oscillations can occur as the network cycles between overload and underload.

This can have two negative consequences in terms of network usefulness and throughput. If the load is too high, a decline in effective network capacity occurs resulting in reduced throughput. This may also cause user dissatisfaction resulting in long term drops in demand. On the other hand, if the load is too low due to overcontrol, the network is not even attempting to carry offered load so throughput is reduced here also.

Determining when and to what degree the flow control should respond are complex questions. Oscillations can even occur when no measurement error exists due to discrepancies caused by time lag. Instantaneous load estimates based on queue length aren't reliable and smoothing load estimates over multiple intervals is not perfect due to the existence of "old" information remaining in the estimate that the system may have already corrected for [Ref. 12: p.113].

Even with flow control existing between parts of the network, too many packets can still be present in parts of the network causing congestion. Long queues can build up at nodes and can eventually result in deadlock. Deadlock is when two or more nodes are trying to send each other packets but, due to full buffers, none of the nodes can accept packets. This problem has been solved in centrally controlled networks but distributed networks must frequently rely upon deadlock detection methods and purge the locked-up traffic. This ends the deadlock but causes destroyed packets. Congestion control is then needed and normally uses some method of network access restriction and deadlock

detection/correction. Flow and congestion control should ideally have all of the attributes shown in Table III It is impossible to obtain perfection in all attributes because of counterbalancing effects but a proper mix is needed to meet the performance objectives of the network manager and users.

TABLE III
FLOW AND CONGESTION CONTROL CRITERIA

-Effective
-Sensitive to traffic patterns
-Fair to all users
-Stable (not wildly oscillating)
-Responsive
-Minimal network overhead
-Robust (able to handle a varity of network conditions)

It is plain to see that congestion and flow control go hand-in-hand and must be integrated together. Routing algorithms also interplay with flow and congestion control resulting in a level of system performance. A lot of literature and even whole conferences have revolved around the topics of flow and congestion control. Readers interested in a more in-depth explanation of methods than can be offered here should read Sevcik [Ref. 13] and Kleinrock [Ref. 14].

D. NETWORK MANAGEMENT

Managing a large communications network involves a myriad of decisions. The decisions occur at every level of use and every network location. Many of the decisions were made at the time of network development and are programmed into network node software. Others were not. Infrequently expected events are often best handled by network operators. Too many factors affecting a decision or even oversight may require operators to make network management decisions on a case by case basis.

Some management schemes are centralized, others are distributed. In a centralized system, at any given time, the controlling element is the responsibility of a single entity or node in the network. In order to ensure reliability, a number of different stations can assume the role of network manager but only one at a time. All others serve as alternate managers.

Distributed control means that network management is divided among many, and perhaps all, entities within a network. Examples are a switch performing independent routing and a switch detecting a fault and performing internal diagnostics. Distributed control tends to be very robust since it doesn't depend on control from a single outside entity. However it does tend to be less than optimum since each entity makes decisions based on a very limited picture of the entire network.

The control may be either passive or active [Ref. 11: p.145]. Active control means the controlling elements are continously evaluating network performance and are taking real-time actions to ensure that the network is operating optimally. Active control allows maximum utility to be derived from network resources but the entire network may be severely affected if the control capability becomes defective.

In passive control, the controlling elements are effectively in the outside of the network, looking in. The network is designed to be self-sufficient and the control tends to be long-term, hence probably not optimal at any one given time.

An effective management scheme normally combines all of the characteristics, that is, it includes centralized and distributed control attributes using both active and passive control methods to manage the network. Network management can be defined as a real-time surveillance and control

activity utilizing techniques to optimize use of network capacity under stress due to traffic overload factors. It employs automatic and manual capabilities which recognize the onset of overloads and respond with traffic control actions within a time range of seconds for automatic controls to minutes for manual actions.

1. Control Functions

Table IV [Ref. 11: p.143] summarizes the network control functions that must be performed. Here again, perfection in all areas is impossible and tradeoffs are necessary in actual networks.

TABLE IV

ROLE OF NETWORK CONTROL IN A PACKET SWITCHED NETWORK

Assure Network Integrity
- Connectivity
- Reliability
- Restoration

Maximize Throughput
- Effective routing strategy
- Effective flow and congestion control

Monitor Abuses and Malfunctions
- Privacy and security
- Prevent misrouting
- Prevent disruption and spoofing
- Data accuracy

Cost Recovery and Billing
- Traffic quantity
- Resource usage
- Delay and precedence data

Planning Data
- Traffic growth
- User sensitivities
- Traffic patterns

2. The Network Monitoring Center

In a typical packet switched network such as the DDN minute-to-minute network operations fall under distributed controls located at the individual switching nodes. A Network Monitoring Center (NMC) is used for exceptions to the norm. In the DDN, the NMC monitors network element status, topology, and performance [Ref. 5: p.70]. When

necessary, an operator can exert control via operator commands sent to "fake" hosts that reside at network nodes. In this way, the commands can be sent via message packets along normal network routing. Supervisory and control data can also be sent on separate lines in an order-wire fashion or in out-of-band signals using portions of the link bandwidth allocated specifically for control.

The DDN NMC can perform numerous functions, some of which are outlined below.

1. Assign network resources: Since resources must be shared, some networks try to be fair while others provide different users privileged access in a priority scheme. This may preempt those currently using resources in order to serve high priority users.
2. Ability to load switching software at nodes remotely from the monitoring centers.
3. Database Management: Includes updating and maintaining static and dynamic information on network elements.
4. Adaptive Presentation: Data can be screened or filtered for special attention and categorize it for individual users or files.
5. Communicate securely to other nodes to avoid spoofing or unintentional network control actions.
6. Establish Thresholds: Distributed controls may handle problems up to a certain threshold but once exceeded, the monitoring center may act. Actions may be in the form of establishing additional diagnostic performance monitoring when an unidentified fault is indicated or even complete fault detection, isolation, and correction.

Monitoring centers of large networks are normally manned 24 hours a day to handle problems when the need arises. Normally several monitoring centers are manned, some as primaries and others as alternates. Mobile monitoring centers are available in the DDN should a crisis occur and network reconstitution be needed due to normal monitoring centers being destroyed.

3. Summary

Packet switched networks are simple in nature but complex to implement. Many factors affect its performance and optimizing one may detrimentally affect another. Parameters must be monitored to detect problems and

appropriate control measures taken in real-time to ensure user satisfaction in a large network.

III. MEASURING NETWORK PERFORMANCE

A. INTRODUCTION

Because networks do not possess infinite speed and bandwidth with zero chance of error and are not built of components with an infinite lifetime, we must monitor network parameters to determine its performance. Present technology allows us to approach these limits but budget constraints and common sense tell us that we can normally survive without perfection.

In determining how well a network is performing, there are two fundamental questions that must be asked; "What should we measure ?" and "How should it be measured ?". Let us look at the second question first.

B. HOW SHOULD WE MEASURE ?

There are three methods for determining the performance of a network [Ref. 15: p. 19] :

1. Mathematical Modeling
2. Simulation
3. Data extraction from the system itself (on-line monitoring).

Mathematical modeling employs theories of queuing and flow by describing certain network characteristics in sets of mathematical equations. This is a complex problem and often assumes away critical parameters. These assumptions are the principle disadvantage with mathematical modeling as too many assumptions impose an unacceptable level of abstraction. Although the validity of these modeling techniques is proven, the methodology is often understood only by the modeler [Ref. 15: p. 19].

The second technique, simulation, is an abstraction of concepts fundamental to the problem. Characteristics of nodes and links and the algorithms that control message flow

are developed into a computer program. The operator, in running the program, can inject messages of various lengths and rates along different routes and observe how the network responds. Node and link failure can be introduced as well as changes to the component characteristics and further network response can be observed.

Mathematical models serve to show the best performance that a network can exhibit while simulation can provide a better "global" view of the network's response in real scenarios. In allowing network designers to observe how network characteristics contribute to performance, intelligent decisions concerning design or control alternatives can be made. These methods help us to tell if a network contractor's specifications will be true when a system is fielded and to see how a network responds to scenarios that we don't want to impose upon an actual operating network.

In a typical network simulation run [Ref. 15: p. 77], a statistics package collects and collates vital data such as:

1. average number of packets at a node
2. average number of packets on a circuit
3. percentage use of a circuit
4. average delay over a route
5. number of packets reaching a destination and time associated with them

Real network performance can be evaluated as well as the impact of changing a buffer size or deleting a link or node. The major problem lies in the validity of the model. With valid models we can gain confidence in a network's ability to meet users' needs, as design by trial and error can be expensive for a major network.

The last means of measuring network performance, on-line monitoring, is the most ideal and is the major focus of this thesis. With data extracted from an operating network, the network manager can fulfill certain objectives such as fault detection, isolation and repair, billing, traffic

administration, resource management, system tuning, mode and link evaluation, as well as network model evaluation [Ref. 16: p. 201]. It also allows us to analyze delay and capacity statistics for developing new control algorithms and predict future demands and user needs, as well as for determining baseline network performance levels as was done in the telephone industry just prior to the divestiture of the Bell System [Ref. 17].

Although on-line monitoring may be the ideal means, it is not without problems. First off , the network must exist. The monitoring system often interacts with the network which adds to the system load and affects the neutral influence an optimal monitor should have. Also since evaluating the entire network is too difficult of a task, statistical sampling is done. This introduces a degree of error depending on sample size and its degree of randomness.

C. WHAT DO WE MEASURE

Network performance parameters have been monitored for a long time in trying to meet the objectives stated previously. Historically, many of the parameters originally developed as performance measures such as grade of service, frequency response curves, and frequency shift were developed for circuit switched, analog systems. These parameters no longer fit the characteristics of modern distributed digital packet-switched networks.

Before determining what to measure, another question must be answered - for what reason are we measuring performance? Although the particular parameters may be similar in any case, how we group them, how we use them, and with what accuracy do we measure them will certainly vary. This concept will be further expanded in the next chapter when a network performance measure is developed with the idea of enhancing the benefits to the end user. Performance parameters generally fall into categories concerning throughput,

which is the usable rate at which data can flow from point to point; delay, which is the time it takes for data to get from point to point; and error performance, which is the degree to which the data arrives at the destination free from error. Another category will be discussed here which is a catch-all category, "other factors". These include jitter, link or node status in terms of availability, queue length, statistical information for later evaluation, and figures relating to measurement error.

1. Throughput

Throughput is the effective rate at which data can be sent and is typically measured in bits per second (bps). It can be analyzed from several viewpoints. Effective throughput for a link is useful so a network manager can determine if a problem lies in a particular area of the network. Effective end-to-end throughput is important to the manager in terms of providing a certain level of performance and is important to the user who needs a minimum level in order to meet his needs.

The maximum throughput of a link is determined by characteristics of the transmission medium, distance, and the signalling technique. Simple media such as twisted pair can provide high data rates but only for short distances due to loss, noise pickup, and signal distortion. As an example, twisted pair will pass about 10 million bps (10 Mbps) for a distance of 100 meters dropping off to about 10 thousand bps (10 Kbps) for a distance of 1200 meters [Ref. 10: p.49]. Twisted pair is thus not too useful for the long haul networks we are interested in. Coaxial cable, due to superior frequency characteristics compared to twisted pair, can support data rates in excess of 100 Mbps but require in-line amplifiers every 1000 meters. This can be very expensive for long haul applications. Optical fibers are becoming more widespread in use for long haul telephone

communications. Such fibers can pass over 2000 Mbps for tens of kilometers, are smaller, lighter in weight, have lower attenuation than wire cables and are unaffected by external electro-magnetic fields. Bell Laboratories has successfully tested a 119 kilometer link at 420 Mbps which required no repeaters [Ref. 18: p.53].

Microwave links are capable of passing several hundred Mbps over line-of-sight paths and are widely used where cable right of ways can't be obtained, and over terrain where cable laying would be difficult. Geosynchronous satellites are widely used for long haul international links and can provide data rates in the range of tens of Mbps but at a cost of about 1/4 of a second in transmission delay to and from the satellite.

These data rates specified are maximum rates of the links. Factors will reduce the effective rate of data transmission. Typical packet lengths are on the order of 1000 bits. Since part of the packet is overhead, actual user data in the packet is reduced. If large packets are used, effective user throughput increases for a constant number of overhead bits per packet.

Because many users may be trying to communicate on the network at a particular time, the effective data rate must be divided amoung the users as their respective packets await transmission. If any prioritization of users is utilized, effective data rate for lower priority users is further reduced. As the number of users increases, so does the network congestion they cause. Network flow control adjusts the the rate at which packets are injected into the network to avoid network overload. This helps to ensure a reasonable data rate for all and avoids the limited data rate that exists when the network is overloaded. If errors occur and packet retransmission is required, effective data rate is lowered, but proper use of error detection/

correction and low-error media can keep this to a minimum. All these factors contribute to the lowering of the effective data rate from the maximum allowed by the medium. All is not lost, for in a network with a rich topology, numerous parallel paths exist between source and destination which contribute enormously to the effective user-to-user throughput.

2. Delay

There are several classifications of delay. Some deal with delay at a node or delay through a link, still others measure end-to-end delay. About the only thing in common with these is the measurement units, seconds. Any delay along the route from source to destination affects the overall delay. A delay occurs when the user first tries to access the network. The network host and/or interface must recognize the user as a valid, retrieve the user's account files, determine the user's intents and then set up connections to destination or necessary resources.

Data delay due to transmission time is on the order of 50 to 100 milliseconds for continental coast-to-coast terrestrial links and connections involving satellite links undergo delays of about 270 milliseconds as the signal goes up to the satellite and back down. Delays occur at each node as the packet is received, processed, stored in a queue, and then forwarded on to its next destination. The number of nodes that a packet must pass through may increase during periods of network congestion and with certain routing schemes, multiplying the delay accordingly.

Factors contributing to delay include processor speeds, congestion and the resulting queueing delays. As packet length increases, the longer it takes for packet reception, processing, and forwarding as well as being subjected to longer delays for a given queue length. This can be further compounded if a priority system is involved

and the user is a low priority user. Although transmission delays are generally considered minimal, under low congestion conditions, it can be a dominant factor if the links include satellite hops. Delay can be an overriding concern for certain user groups, particularly packet switched voice and interactive processing users. It is of less concern to user groups involved in periodic large data file transfers or connectionless services such as electronic mail.

There are several techniques that can be used to measure aspects of network delay [Ref. 19 , 20]. Network delay performance based on actual delay measurement is frequently used. Either a sampling is made of the beginning and ending "time stamps" of existing packets at various nodes or else an artificial packet is generated at a host and sent to some "fake" host residing at a destination switch. This packet is echoed back to the originating host so it can determine the packet's elapsed lifetime. Extensive delay testing is normally performed when a major network component is changed, added, or deleted. A more indirect measure of delay can be made by sampling queue lengths at various nodes and comparing their statistical parameters to some baseline.

3. Error

Error performance standards, or more accurately, freedom from error, have been the subject of much animated discussion amoung several standards organizations including CCITT and the American National Standards Institute (ANSI) [Ref. 21: p.13.6.2]. Three quantities have been mainly used as measures of error performance [Ref. 22: p.2.D.1.2] :

1. Bit Error Rate (BER)
2. Percentage of time where BER does not exceed a given threshold value, BER*,(TI).
3. Percentage of error-free seconds (EFS).

BER is defined, in relation to a given measurement time period, as the ratio of errored bits to total transmitted bits. TI is defined as the percentage of measurements where BER is less than or equal to BER*. Having adopted a measurement period of T(0), the percentages are assessed over some time interval T(1), where T(1) is much greater than T0. EFS is defined in relation to measurement periods of one second and is expressed as the percentage of error-free measurements assessed over a time interval of many seconds. A related parameter is EFdS which is the same as EFS only in this case, the measurement period is one decisecond (1/100 of a second).

Bit error rate is the predominant error measure where errors occur at random times such as on satellite links. It has been found that on terrestrial radio and cable facilities, errors are not statistically independent so EFS has been found to be a more suitable measure in these instances. TI has been seen as a suitable measure in the case of digital telephony where the quality degradation due to errors may be roughly compared to additive noise proportional to a constant BER over the duration of a call. Degradation has been found to be negligible for BER better than 10^{-5}. [Ref. 22: p.2D.1.2].

A fourth error performance measure has been under consideration but has not received much international support which is the percentage of error-free messages (EFM). It is normally defined in relation to a typical data message (packet) size of 1000 to 2000 bits. Some feel that since it is independent of the channel bit rate, it may serve the measurement requirements of data users more accurately than EFS in the case of high channel bit rates (i.e. >1.5 Mbps).

Using a suitable measure of error performance, three network components are typically monitored [Ref. 19:

p. F6.5.3] ; node, trunk, and access link. Node error occurs when a correctly received packet is transferred containing an error. Trunk error has two associated parameters; trunk packet error which is when the CRC check fails on a received packet compared to the total number of packets received, and trunk packet retransmission which is the number of retransmitted packets compared to the total number of packets transmitted. Retransmission rate is typically higher as retransmission can occur for a number of reasons other than CRC errors such as buffer overflow and loss of acknowledgements. Similarly, there are two parts to access link error performance; link frame error rate and link frame retransmission rate.

Another error rate is also typically quoted which is based on undetected packet errors. Continous measurement of this parameter would be difficult for obvious reasons.

Error performance is typically based on error statistics collected from various components of the network. They are typically reported to a controlling station in such a way that high error rate components can be identified. Data applications require a significantly better error performance than voice applications in order to be effective.

4. Other Parameters

In addition to the previously mentioned performance parameters, several others are frequently used. Availability and accessibility are used interchangeability in some of the literature. The term "availability" will be used here and is generally defined as the probability that the resources of the network are available to the user at a particular time [Ref. 19: p. F6.5.3]. Network availability can be broken down into component availabilities, which are the fractions of a selected time interval during which the component is capable of performing their assigned functions. Components

usually monitored are line processors, trunk processors, user access links, and node processors. From these component availabilities, the node-to-node availabilities are calculated. These represent the fraction of time that a path between nodes (from line processor to line processor) is available to set up and operate a virtual circuit or provide datagram service. Usually a simple arithmetic average of the node-to-node availabilities can be used as a network availability index.

Contributions to unavailability may be categorized as short-term or long-term due to the relative impact of long and short-term outages on the user. For example, a voice user subjected to a long term outage may impatiently hang up. Short-term outages disturb both data and voice users due to data corruption but don't usually require circuit restoration action by network operations and maintenance personnel. Multipath fading in radio transmission, resynchronization in time division multiplexing, and automatic switching of redundant equipment are all sources of short term outages. Long term outages consist primarily of catastrophic failures, although certain propagation conditions such as satellite signal fading due to rain may also produce long term outages.

One organization, the CCITT, has defined a circuit to be unavailable when the service has degraded for periods exceeding a time interval (e.g. one second average bit error rate greater than 10^{-4} on a 64kbps voice and data channel that occurs for a period greater than 60 seconds). In this definition, disturbances less than 60 seconds are covered under error performance measures [Ref. 20: p.17.1.3].

Another performance parameter that is often determined for a network is reliability. It is related to availability but deals primarily with the probability of service interruption due to random failures. Reliability can be

expressed in terms of "mean time between failure" (MTBF), whereas availability is more closely related to "mean time to repair" (MTTR). Here again, component reliabilities are determined and combined to form node-to-node reliability measures. This performance measure is normally used in network design and implementation and not used during normal network operation.

Since processors in the path of the packet determine the value of a bit (0 or 1) by sampling, problems can arise due to small variations of the sampling instant from its intended position in time or phase. This is known as jitter. Larger time variations are known as wander or drift. Jitter can occur at any processor along the way. The methods of controlling jitter below a required level are available and used in networks and are very complex.

The last performance parameter to be discussed pertains to synchronous networks and is called "slip". Slip occurs when a momentary loss of synchronization occurs due to errors in the clocks used to control network processors. At high bit rates, a small variation in time measurement can cause slip. The result is usually a bit or packet error causing an occasional click in digital telephony and packet retransmission in the case of data [Ref. 22: p.2D.1.3].

Additional data that is normally monitored and used for management and operational control include queue length which can indicate congestion or component failure, alarm indications for fault troubleshooting and component utilization data.

D. STANDARD NETWORK PERFORMANCE PARAMETERS

New applications for data communication services and the trend toward deregulation and competition in the telecommunications industry has created a need for specifying and measuring the performance of these services. Various organizations have been pursuing a network performance standard

including CCITT in the international scene, numerous foreign national communication agencies, and U.S. organizations including the National Bureau of Standards, the National Telecommunications and Information Agency (NTIA), and the American National Standards Institute (ANSI). Close coordination with industry is maintained in the process as a standard with no followers can hardly be considered a true standard. Historically, network performance standards have been biased toward technical, network-oriented perspectives. Over the past several years, standards organizations in the U.S. have been working together in developing system independent performance parameters and measurement methods using the end-to-end users as the perspective. Standards have started to be promulgated within the federal government in the form of Federal Telecommunication/Federal Information Processing Standards and in industry in the form of American National Standards [Ref. 23: p. F6.6.1].

Two related performance standards have been developed. The first, American National Standard (ANS) X3.102 defines user oriented performance parameters. The second, proposed Federal Standard (FS) 1043 defines companion measurement techniques. It is expected that these standards will promote innovative and fair competition in the data communications industry by providing users with a common denominator for measuring a delivered performance and comparing alternative network services. It is estimated that user savings of over $400 million per year can be realized in the total Federal data communication costs through the promulgation of these standards in the near future [Ref. 23: p.6.6.1].

Figure 3.1 illustrates the twenty-one user-oriented performance parameters defined in ANS X3.102. Performance is expressed relative to three primary communications functions: access, user information transfer, and disengagement. Each of the functions is then considered in relation to the

criteria of speed, accuracy, and reliability. One or more "primary" parameters were defined to express performance relative to each criterion/function pair. Four ancillary parameters are also defined in X3.102 relating to the primary "speed" parameters of each function. They express the average portion of of the parameter's performance time that is attributable to user delays. For example, the parameter "Access Time" can include delays attributable to the user such as system interrogation and user response, as well as delays attributable to the system such as switching, queueing, and transmission delays.

These ancillary parameters have very important uses. They can allow for factoring out user influences on the primary speed parameters to produce user-independent values for characterizing system performance. There are three general applications for the X3.102 performance standards [Ref. 23: p.F6.6.2] :

1. User Requirement Specification: The parameters can be used to specify the communication performance requirements of a particular user. An analyst can assess the impact of communication performance on user processes without presupposing any particular network designs.
2. Service Performance Characterization: The parameters can be used to characterize the end-to-end performance of a system or service. This can provide vendors/suppliers with a single, uniform method of representing performance to all potential users
3. Service Selection: The parameters can be used to compare alternative means of meeting user requirements. Thus they provide the communications manager with a practical method of evaluating service utility.

Proposed Federal Standard 1043 provides an overall approach used to define the measurement methods in X3.102. It is written as a specification for a conceptual data communications measurement system consisting of four major subsystems [Ref. 24: p.31].

The Data Extraction subsystem observes signals transferred across user/system interfaces in real time, determines the performance significance and its time of occurrance. It then outputs this performance information in

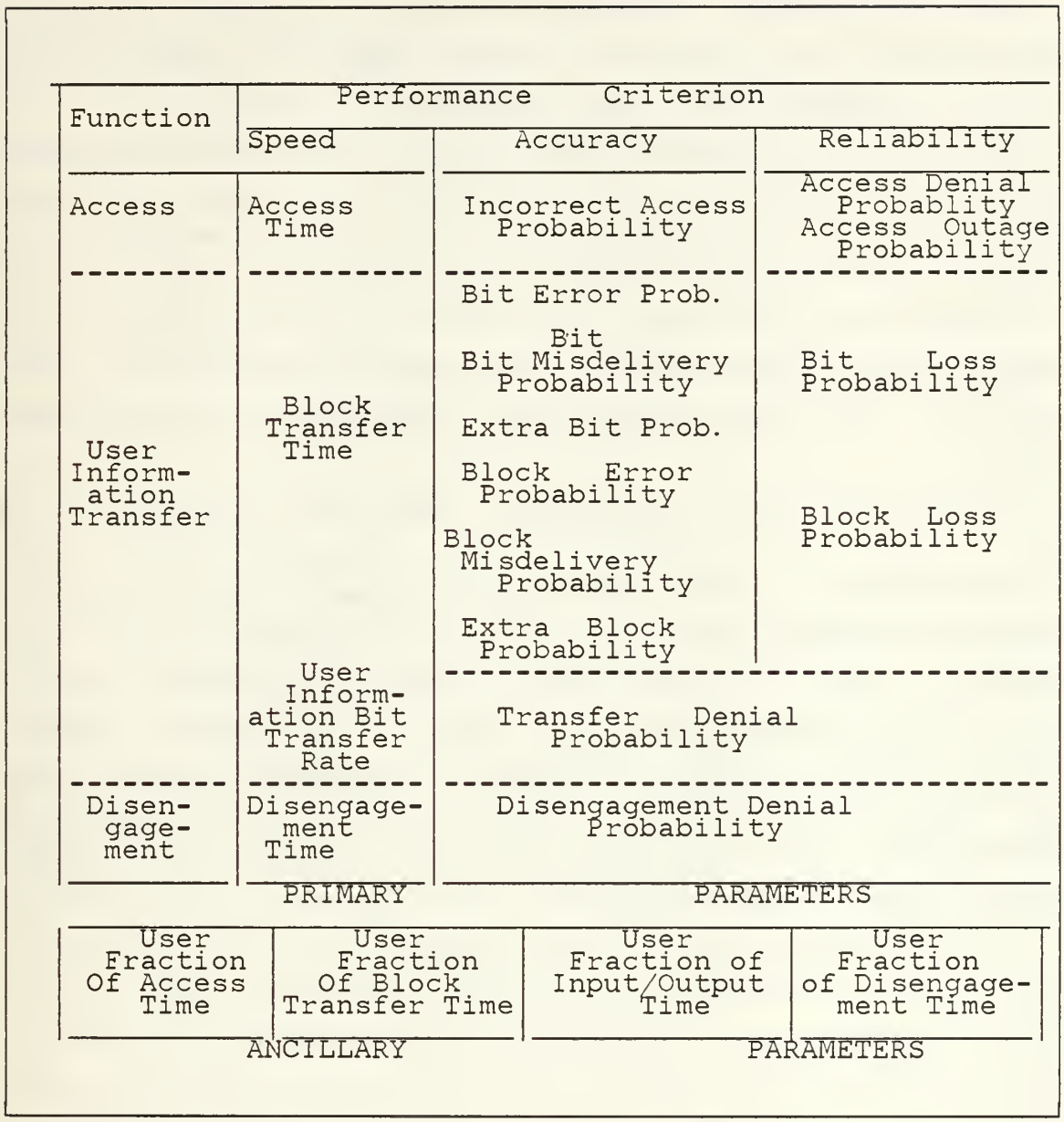

Figure 3.1 Summary of ANSI X3.102 Performance Parameters.

the form of a chronological event history. These events observed at each interface provide the basis for performance parameter calculations. The interface monitors must perform three functions. The Input function detects the signals and determines their time. The Processing function takes these signals and determines the event's significance in terms of

X3.102 performance parameters and the Output function then outputs the nature and time of each significant event in the form of an ASCII record for filing.

The Data File subsystem consists of a set of standard ASCII character files which record the the real-time event histories from the Data Extraction susbsystem. They are available for off-line reduction and analysis. The files are are broken down into four sections; a source overhead information file, a source user file generated by the interface monitor, and destination user overhead and information files.

The Performance Assessment subsystem correlates the ASCII files into standard X3.102 parameter values. The resulting output consists of five pages of data including parameter values, user inputs, and raw parameter data. Each batch of data is associated with a specific source/destination/originating user triplet. In order to characterize the performance of duplex or multipoint transactions and multi-user networks, it is necessary to aggregate performance data from two or more batches to produce composite performance values.

The Statistical Design and Analysis subsystem transforms user-defined measurement precision objectives into appropriate statistical test design and data analysis criteria. These criteria are used to control the data extraction and performance assessment subsystems and produce confidence limits for each measured parameter value. Three major topics are addressed: Qualitative design (i.e. how variables such as time of day, traffic loading, selected user pairs, etc. should be accounted for in arranging measurements); Sample size determination; and lastly, Data analysis (i.e. how to analyze measured data to check pre-test assumptions and to refine measurement precision estimates).

Performance measurement involves a trade-off between benefits and cost. The primary benefit of data communications performance measurements is an improved ability to make decisions about procurement and operation. These benefits are strongly influenced by two factors; accuracy and completeness with which the parameters characterize the service, and the extent to which the values may be compared with values of other services or with specific user requirements. Measurement costs include loss of productive service usage as well as the time and resources required to obtain and structure the performance data. Inevitably, the more accurate and comprehensive measurements cost more than less useful ones.

These two standards were tested using the ARPANET as a test-bed [Ref. 24: p.39]. The tests were conducted over a two month period and the results were felt to be useful by network managers for performance assessment as well as comparing performance to user needs. These standards have already been utilized in a number of trial federal procurements. The standards are continually undergoing improvement and are being pushed for more widespread acceptance.

IV. A USER-ORIENTED PERFORMANCE INDEX

A. BACKGROUND

User oriented performance measures presently under development, such as ANS X3.102 and Federal Standard 1043, indicate a significant departure from typical technical network operation and management oriented performance measures. One must recognize that these user oriented measures are not meant to completely replace the complicated network oriented parameters but rather to supplement them. This is because the user oriented measures are incomplete in terms of providing necessary information for network control, and for fault detection/isolation/correction [Ref. 25: p. 2C.5.1]. These new measures such as X3.102 are proving to be valuable for side by side system comparison, characterization of the performance of a particular system, and for specifying user requirements. It shows great progress in trying to monitor performance from the user's perspective. If one could utilize some of the parameter characterizations from X3.102 and some of the real-time monitoring capabilities of the network oriented monitoring methods, it is foreseeable that an accurate, near real-time indication of network performance could be made available.

Characteristics of packet switched networks are dynamic, so one also expects that the perceived performance of the network is also dynamic. For networks with finite amounts of resources, as the number of users increases and stretches the existing resources, the network manager must either continue to expand the network to maintain the desired level of performance or manage the resources and users better. In exercising user control, managers can either directly control the user or he can provide the user with a means of self-control. Direct control can be in the form of limiting

network access to a certain group of users or limiting all users in the rsources that they can access. Self-control can be in the form of a pricing policy (i.e. charge low rates at what are typically low usage times to encourage users to move away from high usage times). Normally the period of maximum charges is not uniformly congested so user perceived performance will vary. If some dynamic index of network performance can be made available to the user, the user can determine the best times to use the network and receive better performance for a given usage charge. Benefits for the user also include better time and resource usage as now the user can decide the extent of immediate access versus expected performance tradeoffs. Benefits to the network could be in the form of savings on the capital investment for resource expansion by evening out the user demand (hence resource usage) over time. Additional usages of providing the user an indication of network performance could involve usage pricing based on the user's effect on network performance. In other words, charge the user higher prices when performance is poor to encourage him to wait. This would reduce the effect of additional users on the performance of those already using the network.

American National Standard X3.102 does not provide such a real-time user oriented performance measure but does provide some insight as to how to characterize network performance. One of its weaknesses is that it doesn't differentiate between user types. Networks of the future will likely integrate voice, facsimile, interactive computing, electronic mail, and large data file transfer services into a common network. Each of these has different requirements for performance in order to be effective. As network resources get stretched under heavy load, the manager may desire to tradeoff some aspects of performance in order to provide acceptable service to the predominant user type or to a certain segment of the network.

Another weakness of X3.102 is that it is not designed to be dynamic enough for real time measurement. Although the required state-of-art technology needed to provide a real-time performance index for user groups probably exists, it would be very expensive both in capital costs and in impact upon existing network resources.

The parameters of X3.102, although useful, have too many dimensions for real-time use. Twenty-one parameters to evaluate is beyond the comprehension of most users. This chapter will address a method of categorizing user types, performance needs for a user type, and develop a usable index of performance specifically for the user type.

B. PERFORMANCE REQUIREMENTS FOR VARIOUS USER TYPES

Although a large number of ways exist to categorize user groups in an integrated network, groups will be categorized here by traffic type. Three user types will examined for illustrative purposes; voice users, interactive computing users, and large data file transfer users.

Voice transmission, essentially "telephone calls", on a packet switched network are digital representations of the users' voices in packet form. Although once digitized and encoded, they resemble digital data transmissions, but cannot be treated generally like data [Ref. 26: p. 2.2.2]. Conversational voice signals give rise to traffic with moderately high throughput and low variable delay requirements. Information integrity, on the other hand, is not as critical. Voice users typically perceive performance subjectively. Delay beyond about 400 milliseconds becomes objectionable under most circumstances and error rates in excess of about 10^{-5} start to become objectionable. This can vary due to whether the error causes packet loss or just a dropped bit within the packet. Most links can usually provide this level of error performance. Experimental packet voice circuits in present operation normally utilize error

checking for packet overhead (addressing and sequencing) but neglect it in the actual data portion of the packet.

Interactive computing service similarly requires low delay, relatively low throughput, but high data integrity (low error rate) in order to be effective. Large data file transfer users typically desire high throughput and data integrity, but will still be effective with longer delays than voice and interactive computing users.

C. CHARACTERISTICS OF A USER ORIENTED, REAL-TIME PERFORMANCE INDEX

A performance index, in order to be useful, must possess certain characteristics. Although not exhaustive, the following are a basis for the desired characteristics:

1. The index should be understandable and available to the user on a real time basis.
2. The index must give a reasonably good indication of what actual network performance is.
3. It should indicate the degree to which user needs can be fulfilled.
4. It should be derived from actual network conditions and existing network performance measures.
5. It should exhibit reduced dimensionality.
6. It should be responsive to changes in actual network performance.
7. The index should be stable and not subject to wild oscillations.
8. The index should not be swamped by a high value of some performance attribute masking marginal performance in another attribute.
9. Some indication of performance stability would be desireable.

The first three characteristics require that the index be oriented to the user and available to the user either on demand or displayed continuously. The actual display could be in the form of a number on a screen, a meter movement, or even a LED display on a telephone handset. The index must be related to the parameter levels which most directly affect performance as perceived by the end user. Those parameters are related specifically to data rate, end-to-end delay, and service quality.

The fourth characteristic indicates a desireability to not only utilize existing performance parameters of the network, such as X3.102, but also some of the parameters being monitored for network management and operation benefit. This ties actual network performance into a user available index. The fifth characteristic, reduced dimensionality, is vital if the index is to be used in a user session basis. Even X3.102, which is a "user-oriented" performance parameter set is not designed for real-time user evaluation, due in large part to its significant number of parameters. Some method of combining pertinent parameters must be done to provide a one or two dimension index is needed. If properly done, this will ease the real-time decision making process. Although an exact mapping of such parameters into an index is beyond the scope of this thesis, a suitable combination technique will be developed. Any method of combining separate, dissimilar performance measures into a single index involves making trade-offs as to the relative importance of each measure. This thesis will also address a method of specifying these relative weights.

The sixth and seventh characteristics of responsiveness and stability, although working in direct opposition to each other, are still desireable properties of our performance index. A method that does not respond to actual changes in network performance is useless as its value may bear little resemblence to the true level of network performance. In addition, if the user is able to get a subjective "feel" of the network's performance (e.g. noting a longer than usual delay or slow response) and if it is not reflected in reduced performance index level, the user will lose faith in the index as a useful tool and will disregard it. On the other hand, if the index is so responsive that it swings significantly due to performance degradations lasting such a short time period as to be otherwise unnoticed by the user,

the user will again lose faith in the index's value and will disregard it

The eighth characteristic of resistence to swamping is important. For example, a voice user needs moderate throughput, short delay, and relatively low quality. If the network can provide the necessary throughput but the delay is excessive, the index should not change significantly if quality varies from "good" to "excellent". In this case, the overall performance which is poor due to excessive delay should not be masked by large values of quality.

The last characteristic, the measure of stability, can be of benefit to the user particularly during long sessions. This stability measure would be a separately provided measure in addition the composite measure of quality, delay and rate. The stability is more accurately an indication of the rate of change in network performance for a particular user group (voice, file transfer, interactive processor). If the stability measure is provided in terms of direction (+ or - indicating improving or degrading performance) and magnitude (the rate at which the performance is changing due to increased network loading and related problems) the user can make intelligent decisions as to whether the system will meet his needs over the length of his session. This stability measure could be disregarded in the case of most voice sessions if they are only expected to be of short duration (several minutes). This stability measure may also be of benefit to network management and operations forces in determining when to establish measures to reduce access or reconfigure the network in an effort to maintain a usable performance level to those users already in the network.

D. THE MODEL FOR ESTIMATING A NETWORK PERFORMANCE INDEX

Of all of the parameters that characterize network performance, there appears to be three measures that most directly affect the user-perceived performance. These are

the throughput or data rate available to the user, the delay in the information reaching the destination, and the quality of the service provided. These three measures are not independent. Actions taken to change one of the parameters will likely have a noticeable effect on the others. The constraints that this places on our model selection will be explained later. This section will take each of these three measures, describe it, discuss methods of expressing its value, and how the utility to the user varieswith the parameter value. Then the measures will be combined into an index which fulfills the desired characteristics previously described.

1. Throughput (Rate)

Throughput has been defined as the effective rate at which data can move from source to destination. Although one can argue that "more is better", one can also ask "Is twice as much twice as good?". In general, a throughput below some minimum acceptable value provides the user with very little meaningful utility and above some higher value, no additional benefits are gained. As an example, if a network can only provide a user with an effective throughput of 10 bps and his needs are 1200 bps, the network throughput will be of little value. If the network now provides an effective throughput of 800 bps, it will be of more benefit to the user, and even more benefit if the effective throughput is 1300 bps. Is the user any happier if now the network can provide an effective throughput of 2400 bps? Probably not.

A significant factor in determining the necessary data rate is the user type. Interactive processing users have generally low throughput requirements. Voice users integrated into a packet network have a significant data generation rate but due to the bursty nature of human speech (<40% duty cycle) only a moderate throughput is required. Users involved in large data file transfers require a

significantly higher throughput due to the high data rates that the user's terminals/computers are capable of and because the users do not wish their resources tied up waiting to slowly transfer data files when they could probably be better used on another task. If a user must be present at the terminal (which is getting less and less likely as terminals get more sophisticated) the cost of that person's time plus the opportunity cost associated with what he could be doing must also be taken into account.

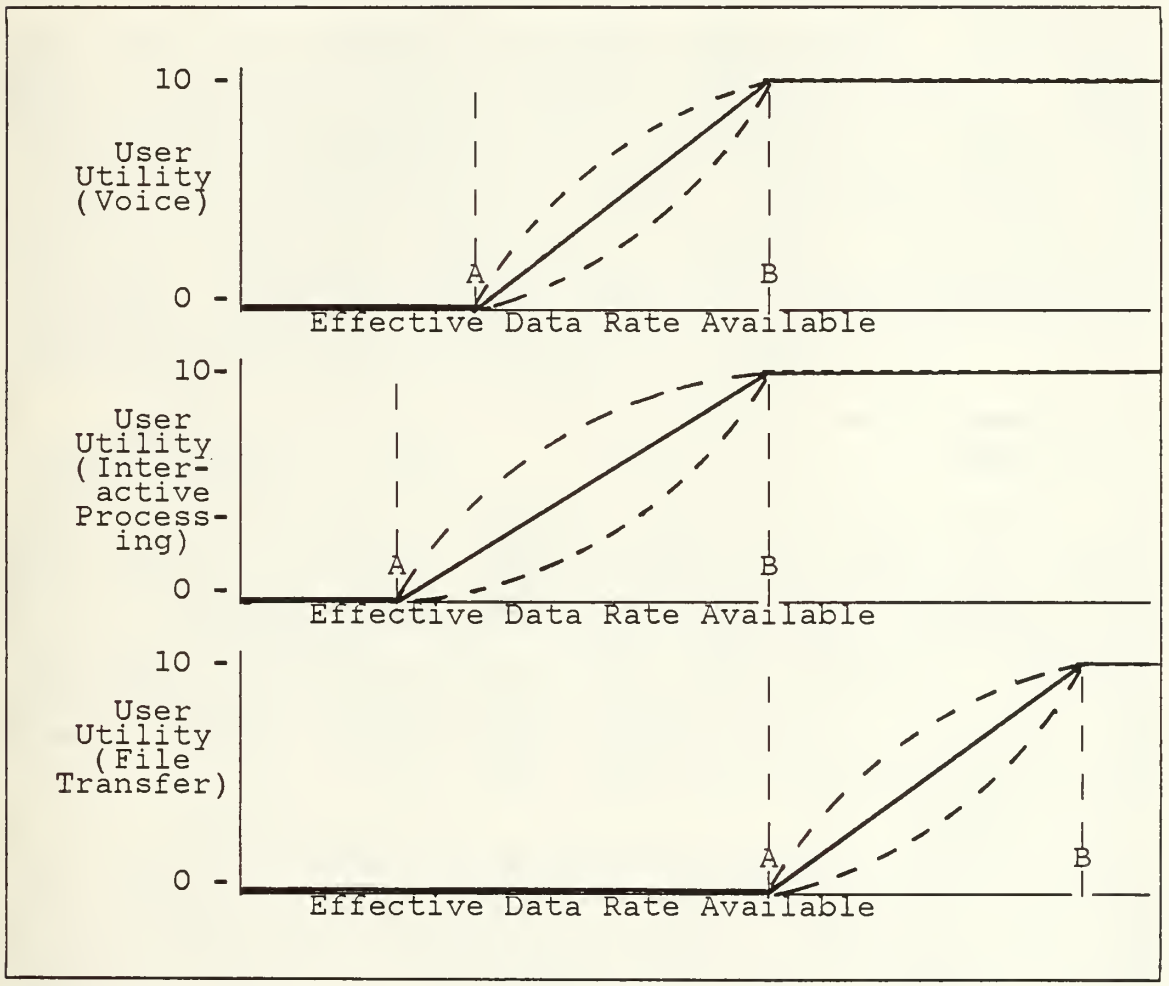

Figure 4.1 Utility Versus Throughput.

The curves shown in Figure 4.1 qualitatively shows the utility or benefit to the user as a function of

increasing the effective data rate available to the user. We expect that the utility would be low (zero) for throughput values less than the minimum acceptable value. The minimum acceptable value is lowest for interactive processing users, then voice, then finally for large data file transfer users. As data rate increases, more utility is derived until some level is reached where no more benefit is derived from increasing the data rate. The utility associated with this maximum level is assigned an arbitrary value (10). The actual values of minimum acceptable rate and the rate associated with maximum user utility (indicated by the lines marked 'A' and 'B') will actually vary from user to user based on his own personal needs and equipment capabilities. Some determination of a nominal or average value would have be required for actual network implementation. Linear, concave upward, and concave downward curves are shown for illustrative purposes. In later utility versus parameter figures, only the linear relationship will be shown. The exact shape of the curve between 'A' and 'B' is unknown. It could actually be a combination of linear, concave upward, and concave downward segments as seen in Figure 4.2 In order to meet our required characteristic of being derived from actual network conditions and performance measures, it is the author's opinion that the "user information bit transfer rate" of X3.102 would be the best parameter to use as the rate measure on the horizontal axis.

2. Delay

Delay is the amount of time it takes for data, once inserted into the network, to arrive at the destination. This is the information that "block transfer time" of X3.102 provides. Other delays may be important to the user, particularly if sequential communication sessions with different destinations is desired. An example would be a voice user wishing to place phone calls to several different people,

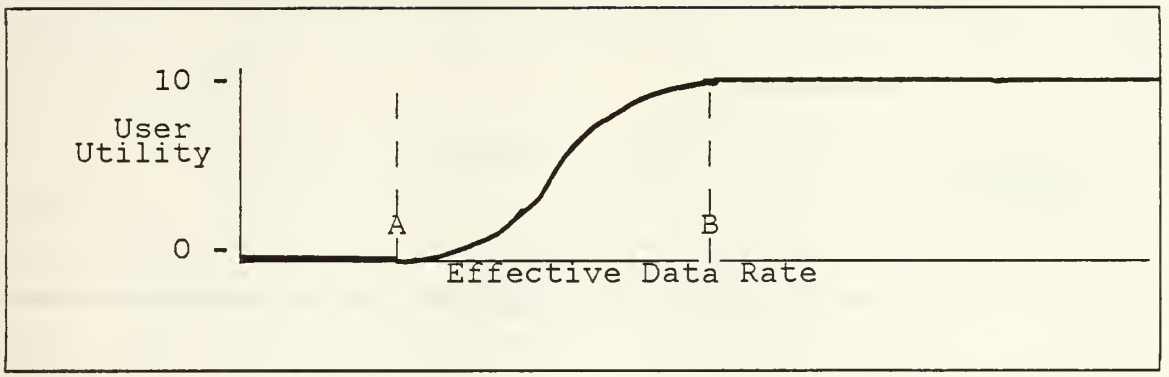

Figure 4.2 Possible Utility Profile.

one right after another. The additional times of interest then include "access time" and "disengagement time". These are the times associated with connection to a destination and the time required to free up the allocated resources and user equipments once a session is complete allowing the next connection to be made. These times could be combined to form a composite user-to-user delay in the form:

$$D = a*T1 + b*T2 + c*T3 \qquad \text{(eqn 4.1)}$$

where 'D' is the user-to-user delay, T1 is the access time, T2 is the block transfer time, and T3 is the disengagement time. The constants, a, b, and c are weighting constants associated with the relative importance of T1, T2, and T3. These relative weights will probably vary for each user group. In the case of voice, the block transfer time takes on great importance so 'b' would have a large value compared to 'a' and 'c'. One can rationalize the relative importance of a, b, and c for the various cases but additional research would be needed to determine the exact values of these weighting factors

Once the composite delay formulation is accomplished, one can then qualitatively plot user utility versus delay for the various user groups as shown in Figure 4.3

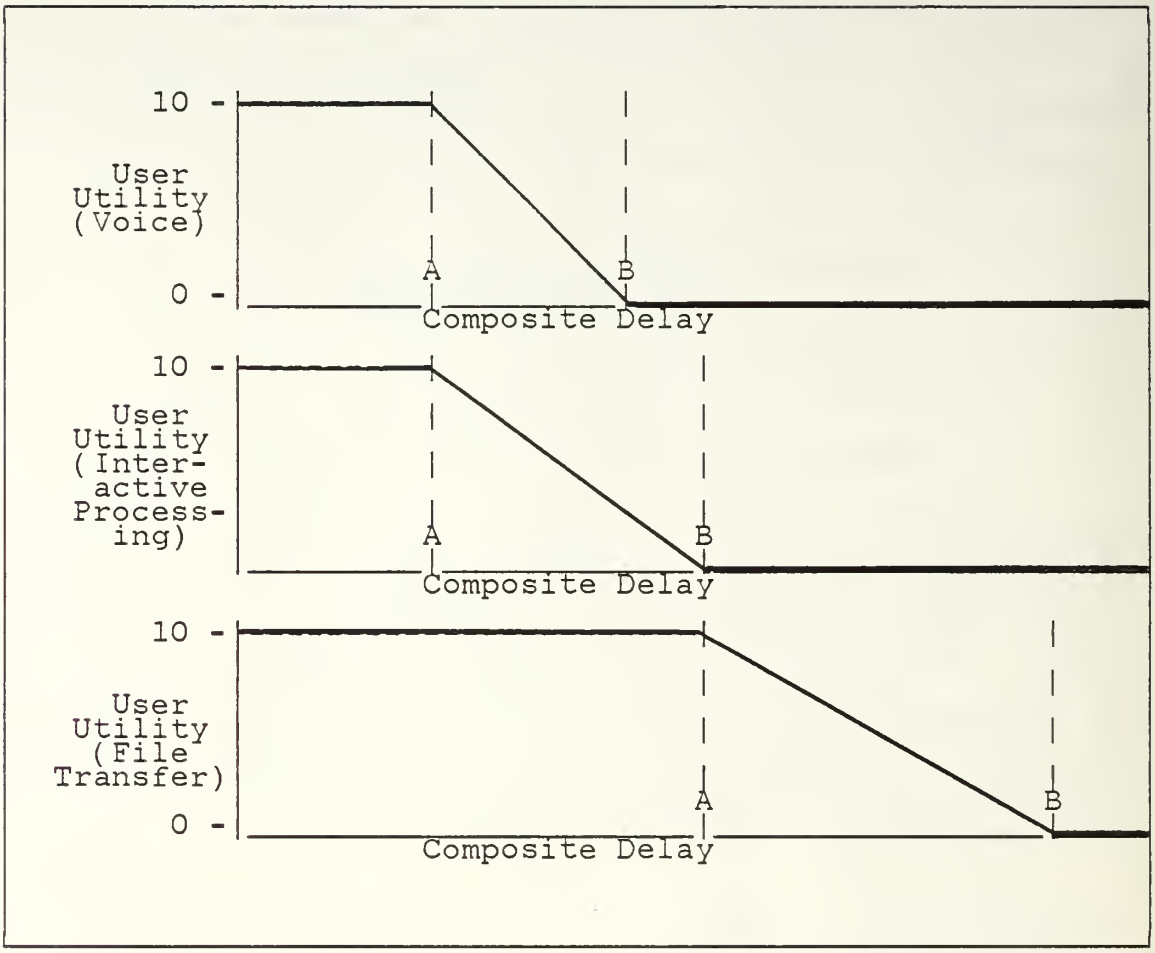

Figure 4.3 Utility Versus Delay.

In this case, maximum utility (arbitrarily set to 10), is derived until a minimally noticeable delay is reached at line 'A'. The derived utility then drops off until a utility value of zero is reached at the maximum acceptable delay at line 'B'. Above this value, the user finds the delay totally unacceptable. It is believed that the desired delay for voice and interactive processing would be relatively short compared to that of large data file transfer. The "window" of acceptable delay is also expected to be narrower in the case of voice and interactive processing than that expected for file transfer. Here again, the exact shape of the curve (linear, concave upward or

downward, or a combination) is not known and would have to be determined.

3. Quality

The quality of service that a network provides is a function of the of the accuracy with which the system reproduces the source information at the destination as well as the reliability by which the network operates. These parameters fall under the accuracy and reliability criteria expressed in X3.102, where accuracy deals mostly with data and address error probabilities and reliability deals with the ability of network resources to remain on line and the availability of network nodes and links to maintain connectivity to those resources. In a homogeneous, integrated network, we expect that the reliability aspect of quality is equally important to all user groups. In cases where this is not true, reliability would have to be weighted accordingly.

Although a functional mapping of reliability and accuracy into "quality" will not be attempted here, a qualitative relationship will shown. Figure 4.4 shows this expected relationship between user-derived utility versus quality of service.

Because of the interpretive ability of the human mind, accuracy (quality) is not as important to voice users than it is to interactive and data file transfer users. This is seen in Figure 4.4 in the case of voice having a lower minimum acceptable and lower maximum necessary quality in order to derive a maximum utility value of 10. Again the exact shape of the curve would have to be determined. Although some differences may exist between the quality needs of some interactive processors and large data file transfer users, they both require generally high accuracy and thus were treated the same.

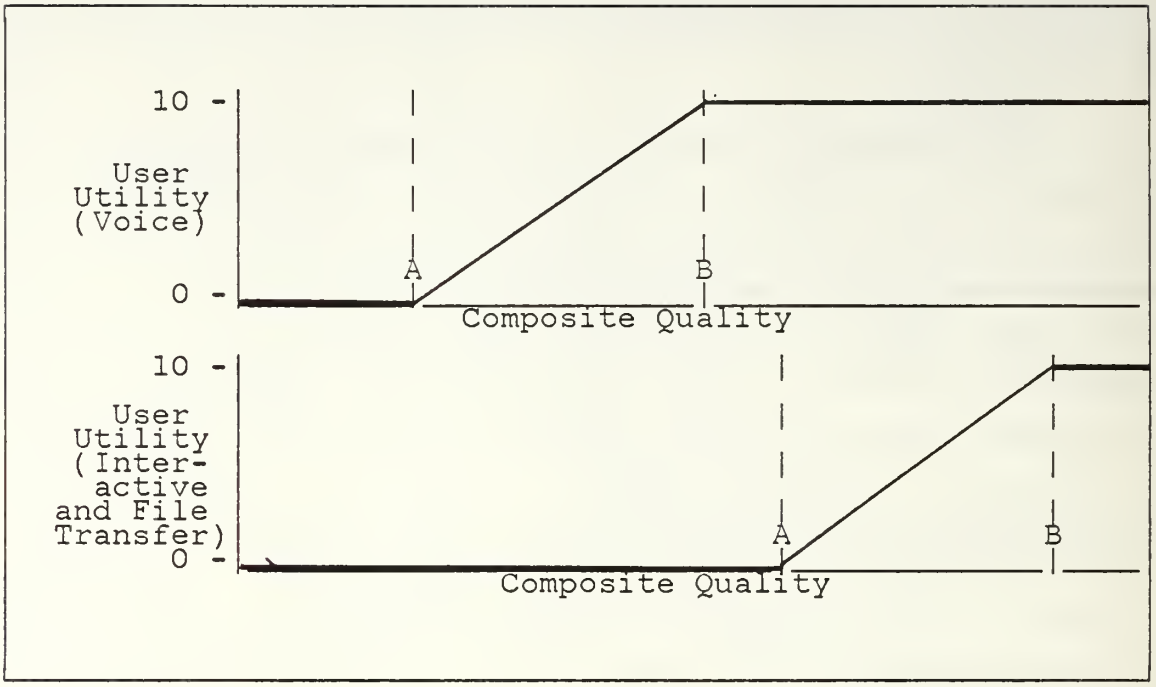

Figure 4.4 Utility Versus Quality.

4. A Composite Performance Index

In specifying user-perceived performance in terms of the delay, rate, and quality, the dimensionality of performance measures has been significantly reduced. In order to meet another of our desired characteristics, we will need to further reduce these three dimensions into a single composite performance index.

Several methods of combining performance measures into a single index exist. One of these methods has already been demonstrated in our discussion of delay where a summation was used. The composite function for performance (P) would then become:

$$P = A * Delay + B * Rate + C * Quality \qquad \text{(eqn 4.2)}$$

where Delay, Rate, and Quality are the utility values taken from the appropriate curves and A, B, and C are their

respective weighting factors expressing the relative importance of the three measures. This is a simple method but does not meet the required characteristic of the index not being swamped by a high value of a single parameter. As an example, if A, B, and C are assumed to be equal to 1, if delay and rate have unacceptable values (0) and quality has a maximum value of 10, performance (P) is equal to 10. This would be the same performance index value in the case where all three measures were 3 1/3. The additive model is also disregarded here because delay, rate, and quality are dependent on each other. Additive models assume parameter independence [Ref. 27: p. 320]. The actual perceived performance would obviously be greater in this latter case.

Another method of reducing the dimensionality is by just using the worst of the three measures as the value of the composite index. As an example, if quality and delay have values of 6 and 7 respectively and rate has a value of 3, then the index has a value of 3. If quality and delay both go up to 10, the index still remains at 3. Even though user perceived performance has gone up, the index has not which does not meet our desired characteristic of responsiveness.

Yet another method of combining the parameters exists, the product of the terms:

$$P = D * R * Q \qquad \text{(eqn 4.3)}$$

If any of the parameters is less than than the minimum acceptable value required, total performance is 0. This equates to the user receiving unsatisfactory service from the network. As long as all the values are not zero, as any of the values goes up, the index goes up and no swamping exists. One problem that arises is that this index is somewhat over-responsive (i.e. if all parameters double, the

performance index value goes up by a factor of eight). A minor modification of this formula solves this:

$$P = (D * R * Q)^{1/3} \qquad \text{(eqn 4.4)}$$

Taking the cube root of the product restores a degree of "linearity" to the index. Taking the 'N'th root of a product, where 'N' is the number of variables in the product, restores a first order characteristic as all the variables are changed. In this case, all the parameters making up the index all have equal weight (delay, rate, and quality all have a maximum utility value of 10). Relative weighting factors can be introduced simply by adjusting the maximum value of utility in the curves relating utility value versus performance parameter value. In implementing this weighting, one should try to ensure that the product of the maximum utility values remain the same for each of the user groups.

This formulation appears to meet all of the desired characteristics dealing with parameter and index values. A multiplicative model form is most appropriate when the parameters are not independent, that is, when one assumes that the variables act conjunctively [Ref. 27: p.321]. Those characteristics that are more strongly dependent on the measurement technique and parameter value computation are beyond the intended scope of this thesis but would require evaluation before implementing such a performance index.

5. Stability

Depending on the nature of the network, its users, and the methods of network control, network performance can have great variability in the terms previously discussed. Periods of poor performance can be tolerated as long as they only last for a short period of time and do not occur very

frequently. If performance degrades during a session lasting only a few minutes, and remains degraded, user satisfaction will also deteriorate. If the network performance cannot be maintained reasonably stable during a user session (unlikely when some sessions may last an appreciable time), then some indication of the network performance stability would be useful.

This determination of stability would be difficult to perform. It would involve short-term, real-time predicting of network usage trends. An instaneous derivative of the performance of the user group performance index would probably fluctuate too rapidly to be of use, so a damped or time averaged measure of the rate of change of performance over time becomes a better alternative.

Short term sessions such as typical phone calls lasting only several minutes may not benefit from such a stability measure but longer sessions, particularly interactive processing and data file transfers requiring periodic operator action could particularly benefit from such a stability measure. As mentioned before, the objective would be to help the user plan network usage to fit his needs and time constraints so he can obtain the maximum benefit fromthe resources he utilizes in his network operations.

V. THE FUTURE

Nodal performance is an abstraction that most users and providers can loosely define but when forced to quantify it in universal terms, they often fall short and have strong disagreements. This is a key reason why standards are so long in development. The transitioning from purely network provider oriented measures to a situation where both provider oriented and user oriented measures has not been easy and is far from complete. The need for these measures to be separate but connected arises from a fundamental difference in the ways users and providers view telecommunications networks. To the typical user, the network is a "black box" that moves information from one place to another. The user is not concerned how the movement is accomplished, whether it is by circuit switched telephone lines or by satellite links or by terrestrial packet switched networks. He is also not concerned with aspects of the network's internal design as long as his needs in terms terms of quality and speed are met.

Users need performance measures for two primary reasons: [Ref. 28: p. 30.6.3] (1) to assess the effects of network imperfections on user operations; (2) to perform cost/benefit tradeoffs among networks, service alternatives and time usage. From the users' point of view then, performance is best expressed by parameters which focus on user-perceivable effects rather than their causes within the networkand do not depend on, in their detailed definitions, assumptions about the network's internal design or protocols. Such measures may be characterized as network independent.

To a network service provider, the network is a collection of connected subsystems, each playing a specific role

in realizing an end-to-end service. Unlike the user, the provider is vitally concerned with the technology and design of the individual subsystems, their specific performance capabilities, and the interaction between them. Because the provider is responsible to all users, he must concern himself with resource utilization, network operations, and facility maintenance issues that aren't of direct concern to any specific user. His interest in performance centers on relating end-to-end performance objectives to subsystems. From the provider's point of view, the performance parameters should focus attention on specific network parts and protocols being used and should be useful in identifying causes of user-oriented performance effects. In this regard, the provider-oriented parameters should be network specific.

To be useful, the user-oriented and provider-oriented measures should be relatable. American National Standard X3.102 and its companion Federal Standard 1043 meet these needs except they do not provide the timely information needed for cost/benefit tradeoffs of network service time usage. This is where a real-time network performance index would be useful.

In determining real-time user-oriented performance measures, enormous amounts of data must be gathered, sorted, evaluated, and outputed in real-time. Equipment exists to measure the required parameters on a single connection but to test all possible connections in a large network would require a tremendous effort. Even to sample a statistically significant fraction of all connections presently would be expensive to do as frequently as we desire.

One solution to obtaining the timely information on network performance at reasonable cost is not to make performance measurements on end-to-end connections but rather on the components that make up the connections such as loops, switches, and transmission facilities [Ref. 25:

p. 2C. 5. 3]. These are the provider-oriented performance measures discussed previously. The number of possible components will be much smaller than the number of ways that they can be put together. Once the information is available on the performance of each component making a connection, the component parameters can be combined and translated into parameters used to characterize the end-to-end performance. These are then termed user oriented performance measures.

The resources involved in continously monitoring all of the components in a network is still a large task and is not yet done on a continous real-time basis for a reasonable cost. A possible solution lies in a technique of network performance monitoring that has been recently developed known as "Perturbation Analysis" [Ref. 29].

Perturbation analysis is an emerging technique for performance evaluation and on-line network control based on actual network data. Given some performance measure, P, and associated parameters (e.g. quality, delay, and rate), it is possible to estimate the rate of change of the measure from observations even though the exact functional relationships may not be known. The application of perturbation analysis involves a "black box" connected to the system (network). This is an on-line processor whose function is to observe the actual network state, evaluate the perturbed state resulting from small changes in parameter values and ultimately provide the perturbed performance measure values.

Although the exact formulation of algorithms is far beyond the scope of this thesis, perturbation analysis consists of three steps. The first is perturbation generation which involves the mapping of a parameter perturbation into a perturbation in a service completion event. Then perturbation propagation deals with how the perturbation moves through and affects the entire network, and finally the third step is the calculation of the effect on the performance measure.

One assumption that present perturbation analysis is constrained by is that perturbations must be sufficiently small so that the order of events remains unchanged at each node or process. Violation of this assumption can result in considerable degradation of perturbation analysis performance. This severely limits the applicability of it to real networks. Research is currently being done dealing with improving perturbation analysis accuracy and gaining a better understanding of its limitations [Ref. 29: p.13.5.4].

A. CONCLUSIONS

Modern packet switched networks have undergone significant changes since their birth in the 1960's. They have evolved into almost "universal translators" converting information from one source into a digestible form for use at a significantly different destination, supplying speed, coding, and protocol conversion as necessary [Ref. 2: p.1311]. Nearly all of the packet switched networks have dealt strictly with data but the time when voice and other users will become integrated is nearing. Voice, in an integrated network will present problems because it must be handled differently to be of use to the user.

Integration of voice does provide some interesting possibilities. It is a highly variable rate information source possibly permitting great network flexibility under dynamic traffic conditions. This is due significantly to the natural statistical multiplexing characteristics of packet switched networks.

As future networks evolve, it is clear that networks will be integrated. Even though significant cost/benefit advantages have been projected for totally packet switched networks due to improved efficiency in resource sharing, and system flexibility and modularity [Ref. 30] the world where all networks are packet switched networks is very unlikely. Due to tremendous capital investments in existing resources

and that needed for complete packet implementation, networks will likely be hybrid networks. These hybrid networks, where aspects of circuit and packet switching are used in some parts or all of the network, point out the need for user-oriented network independent performance measures such as ANS X3.102. This type of measures have shown great usefulness for service selection of one network over another, specifying user requirements in performance terms, and in characterizing the service performance of a network presently being used. Standardized performance measures have demonstrated the potential to save several hundred million dollars annually.

As networks such as the public telephone network evolved, user habits have caused degrees of congestion where resources are stretched so thin, at times, that they are useless to those wishing to access the network. Some of the problems have been solved by resource expansion and a lot has been solved by management actions to spread the demand more evenly over time. Actions such as reducing usage charges for off-peak hours have worked well but a large amount of demand variability still exists during peak priced hours. A similar situation can be expected to occur in packet switched networks as they become more widely used. If the variability in congestion (and the related performance degradation) could be quantified and made available to the user, a means of user self-control exists. This performance index could prove valuable to users in trying to gain more benefit from their own resources and available time.

This real-time index must have certain inherent characteristics as outlined in the previous chapter of this thesis. Some means of combining the most relevant performance parameters that the user is concerned with (rate, delay, and quality) is needed for ease of use.

In integrated networks, one can treat all user data as the same but it has been shown that user needs are not uniform. Some data types require significantly different levels of performance attributes in order to have adequate service. A real-time performance index must acknowledge that fact and model itself more closely to user needs.

In addition to the users benefitting from such a real-time performance index, the network manager/provider can benefit as well. In his actions of controlling and tuning the network, the manager would be able to see how his actions affect user-perceived performance. Future networks will be flexible in packet sizing, flow and congestion control, and routing algorithms. Actions taken to benefit one parameter may have detrimental affects on another and unless the manager can quantitatively see how his actions are affecting the users, he can't be sure his decisions are the best possible. The manager may have the capability to fine-tune the entire network for a predominant or highest priority user type. He may be able to fine-tune a specific portion of the network in the same fashion. The capabilities of the future are unknown today.

In providing a new network or in developing new service capabilities, the network designer must always keep the user in mind. New systems must strive to to supply performance better than or equal to that which may be presently be provided. It is also here where user oriented performance parameters may be of benefit. Modelling network designs to meet user needs will help ensure network systems meet those needs without needless expense for performance improvements which are not perceivable to the user.

B. AREAS FOR FURTHER RESEARCH

As in many instances research, the research often results in more questions than answers. These questions result in more areas of research to be explored (or at least

considered) prior to implementation of an idea. It is obvious to this author that, even though there are benefits to be gained, the implementation of a real-time user-oriented performance measure is a long way off.

One of the foundations of this idea is that the peak demand created by the users is greater than what the network resources can supply, resulting in congestion. If network capabilities grow faster than demand such as in the case of the DDN where plans exist to expand the backbone capacity from 56 kbps to 1.544 Mbps, congestion may be alleviated until demand catches up [Ref. 2: p.1311]. Resource growth is often in bursts while demand usually increases inexorably.

It has been shown that networks usually require both management oriented and user oriented performance measures. The parameter mapping of the associated parameters into the units related to quality, delay, and data rate needs to be explored. Questions exist in the area of determining, to what degree network usage will be effected by knowing how well the network is performing. This degree of man-machine can be speculated about but a realistic determination of how this knowledge affects user behavior modification and demand smoothing, must be explored. Since each user group is expected to have a different utility versus parameter value profile, one can expect that each user, in his subjective determination of performance, has a different profile. How can these be combined to get a usable utility profile for a group?

Mention was made of using this index as a possible modifier for usage charging. This charging could be based on resource usage, contribution to network congestion, or by user type. What, if any, would be the best way to implement this concept? Since such a real-time index would involve a monetary cost for implementation, at what cost/benefit level does its introduction show usefulness.

The last point concerning the index deals with the delay parameter associated with voice. Present networks use satellite links on some long haul calls. This delay can be annoying but people are accepting it as "the way that it is". If delay takes on less importance to voice users, than can we treat voice like any other data or must we invent another parameter such as "delay variability" to characterize users' needs.

Additional areas needing further research also lie with future networking problems related to integrated networks. As the speeds of these networks increase due to better transmission media and improved hardware/software, more packets will exist en-route between sources and destinations. A determination of whether existing methods of flow control, routing, or acknowledgement techniques will be adequate or if new ones will be needed must be made.

If, as it is suggested, that voice integrated on a network must treated differently due to the subjective evaluation that takes place, how will it be handled? Will it be prioritized with other types of data sent during silent periods or will it separated and handled in a subnetwork? Data compression techniques exist which provide the same information as the original data but incorporating fewer bits. How can this be applied to voice data communications? The question of what to do with lost voice packets also arises. Voice packets not only need to be sequenced but also need some sort of time stamping so that playback at the destination takes on a natural sound. Some network experiments have substituted silence for the lost packet, others have used a tone indicating lost speech. Both of these have unnatural sounding qualities. A better substitute used in other experiments have used repetition of the last correctly received packet to replace lost packets. All of these may have serious problems in secure voice applications where

characteristics of previously received data may affect the decoding of present data. Most of these experiments were conducted on the ARPANET [Ref. 31: p.27.3.3]. These experiments, using 100 to 200 millisecond long packets have demonstrated good quality. Selecting packet length involves a tradeoff. Small packet size causes the percentage of total bits devoted to overhead to increase, degrading efficiency. Long packet size tends to raise the minimum delay due to the time it takes to packetize the data at origination. Experiments with voice , with its extended bursty nature requiring real-time delivery, have shown that these same requirements may also fit military applications involving data retrieval of datafrom remote sensors.

Lastly, in network design and implementation, performance must be equal to or better than that provided by present systems. Work needs to be done to chracterize the performance of the present systems in terms of the user oriented performance measures for this goal to be met.

LIST OF REFERENCES

1. Whitehouse, G.E., Fagan, J.J., Extensions of Stochastic Theory to Facilitate the Development of MOE's for Communication Systems, Industrial Engineering Dept., Lehigh University, Bethlehem, PA, 1977.

2. Roberts, L.G., "The Evolution of Packet Switching", Proceedings of the IEEE, Vol 66: No. 11, November, 1978.

3. Baran, P., On Distributed Communications, Vol. 1-11, Rand Corporation Research Documents, August 1984.

4. Carlucci, F.C., Deputy Secretary of Defense Memorandum. Subject: AUTODIN II Termination, 2 April 1982.

5. Defense Communication Agency , Defense Data Network Program Plan, January 1982.

6. Hagouel, Jacob, "Source Routing and a Distributed Algorithm to Implement It", IEEE Infocom '83, Institute of Electronics and Electrical Engineers, 1983.

7. Caldwell, M., Improvements in Routing for Packet Switched Networks , Ph.D. Dissertation in Electrical Engineering, George Washington University, 1975.

8. Perlman, R., "Fault Tolerant Broadcast of Routing Information", IEEE Infocom 1983, IEEE Computer Society Press 1983.

9. Lo, S.C., and others, "A Distributed Database Design for a Communications Network Control System", American Federation of Information Processing Societies (AFIPS) Conference Proceedings, 1983.

10. Stallings, William, Local Networks, An Introduction, MacMillan Publishing Co. 1984.

11. Rosner, Roy D., Packet Switching: Tomorrow's Concept Today, Wadsworth, Inc. 1982.

12. Fernow, J.P., El-Sayed, M.L., "Stability of Adaptive Congestion Controls in Packet Networks", IEEE Infocom 1983, IEEE Computer Society Press, 1983.

13. Sevcik, P.J., Mayersohn, J., "Evaluation Criteria for Congestion Control Techniques", Proceedings of the 15th Hawaii International Conference on Systems Sciences, Vol. I, 1982.

14. Kleinrock, L., Gerla, M., "Flow Control: A Comparitive Survey", IEEE Transactions on Communications, Vol. Com-28, No.4, April 1980.

15. Jennings, S.C., Hartel R.J., Petri-Net Simulations of Communications Networks, M.S. Thesis Computer Science and Systems Technology (C3), Naval Postgraduate School, March 1980

16. Brendel,P.J., Frank,R.L., "Panel Discussion: Impact of New Sevices and Systems on Network Operations and Maintenance", IEEE Global Telecommunications Conference Globecom '85, IEEE, 1985

17. Cary, and others, "1982/83 End Office Connection Study: Analog Voice and Voiceband Data Transmission Performance Characteristics of the Public Switched Network", AT&T Bell Laboratories Technical Journal, Vol 63: No. 9, November 1984.

18. Stallings, William., Data and Computer Communications, MacMillan Publishing Co., 1985.

19. Unsoy, Mehmet, "Performance Monitoring and Evolution of the DataPac Network", National Telecommunications Conference Proceedings. 1981 IEEE 1981. Page F6.5.4 .

20. Smith, D.R., Cybrowski,W.J., "Performance Standards for Military Long-Haul Digital Transmission Design", IEEE Global Telecommunications Conference, GLOBECOM '85, IEEE 1985. Page 17.1.7 .

21. Clarke, P.G., " Status of CCITT Studies on the Overall Performance of Public Data Networks", IEEE Global Telecommunications Conference GLOBECOM '83, IEEE 1983.

22. Decina, M., Julio, U., "Performance of Integrated Digital Networks: International Standards", IEEE Communications Conference Proceedings 1982, IEEE 1982.

23. Seitz, N., and others, "Data Communication Performance Measurement - A Proposed Federal Standard", IEEE National Telecommunications Conference: NTC '81., IEEE 1981.

24. Seitz, N.B., and others, "User Performance Measurements on the ARPANET", IEEE Communications Magazine, Vol. 21, No. 5, August 1983.

25. Le, N.H., Newcombe,E.A., "Data Network Performance: Specification and Measurement"; IEEE Conference on Communications 1982, IEEE 1982.

26. Le, N.H., Gruber, J.G., "Performance Requirements for Integrated Voice Data Networks", IEEE Global Telecommunications Conference GLOBECOM '83, IEEE 1983.

27. de Neufille, R., Stafford, J.H., Systems Analysis for Engineers and Managers, McGraw Hill Book Co. 1971.

28. Paolucci, R., Seitz, N.B.,"A General Framework for Describing Quality of Service and Network Performance Standards in Digital Networks", IEEE Global Telecommunications Conference Globecom '85, IEEE 1985.

29. Cassandras, C.G., Strickland, S.G., "Perturbation Analysis Techniques for Communication Networks", IEEE Global Telecommunications Conference Globecom '85, IEEE 1985.

30. Gitman, I., Frank, H., "Economic Analysis of Integrated Voice and Data Networks: A Case Study", Proceedings of IEEE, Vol. 66, November 1978.

31. O'Neill, J., Bernet,M., Gan, D., "Secure Packet Voice Integration in a Generic Packet Switched Network", IEEE Global Telecommunications Conference Globecom '85 IEEE 1985.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, California 93943	2
3.	Dr. J.W. LaPatra, Code 54Lp Department of Administrative Sciences Naval Postgraduate School Monterey, California 93943	3
4.	Lieutenant Mark E. Speck 1051 Halsey Drive Monterey, California 93940	2
5.	Department Chairman, Code 54 Department of Administrative Sciences Naval Postgraduate School Monterey, California 93943	1
6.	Mr. Ken Boheim NCS/PP Eighth Street and South Courthouse Road Arlington, Virginia 22204	1
7.	Mr. Edward M. Cain NCS/PP Eighth Street and South Courthouse Road Arlington, Virginia 22204	1
8.	Dr. Bruce Barrow NCS/PP Eighth Street and South Courthouse Road Arlington, Virginia 22204	1
9.	Colonel William Schooler NCS/EP Eighth Street and South Courthouse Road Arlington, Virginia 22204	1
10.	Mr. Norman Douglas NCS/EP Eighth Street and South Courthouse Road Arlington, Virginia 22204	1
11.	Lieutenant Colonel Tom Cindric (JDSSC) Defense Communications Agency Code 62 Washington, D.C. 20305	1
12.	Professor Michael Spencer, Code 54Sp Department of Administrative Sciences Naval Postgraduate School Monterey, California 93943	1

217592 82

Thesis
S6672 Speck
c.1 A user-oriented per-
formance index for
pocket switched net-
works.

1 MAY 89 33632
1 MAY 89 33632
19 OCT 89 33632

217592

Thesis
S6672 Speck
c.1 A user-oriented per-
formance index for
pocket switched net-
works.

thesS6672
A user-oriented performance index for po
3 2768 000 66057 5
DUDLEY KNOX LIBRARY